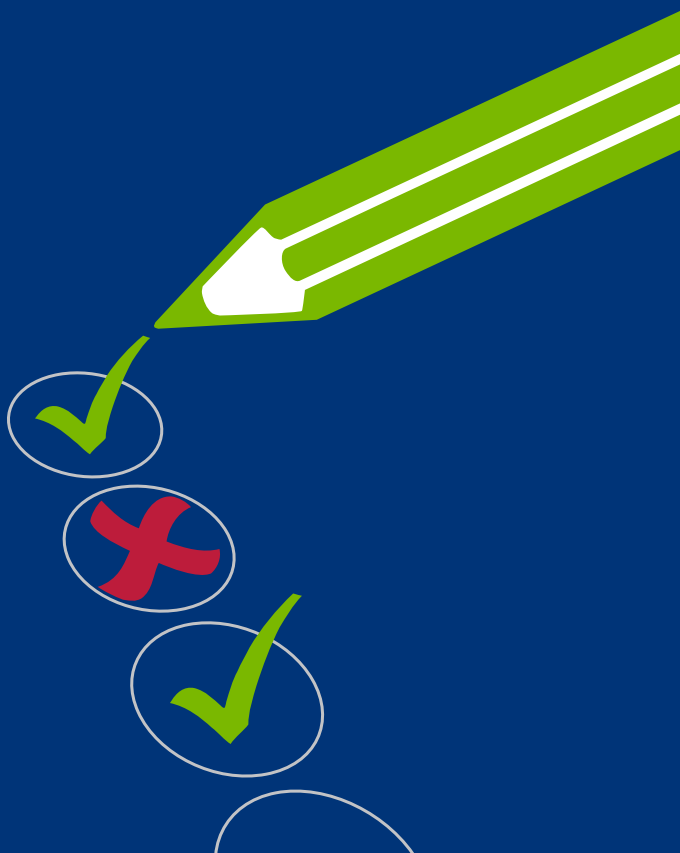
Quality Control
using Semantic Technologies in Digital Libraries

Sascha Tönnies — Dissertation — 2012



Quality Control using Semantic Technologies in Digital Libraries

Von der Carl-Friedrich-Gauß-Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von Sascha Tönnies
geboren am 24. Mai 1978
in Hannover

Eingereicht am:
01.08.2012

Disputation am:
03.12.2012

Referent:
Prof. Dr. Wolf-Tilo Balke, Technische Universität Braunschweig

Korreferent:
Prof. Dr. techn. Wolfgang Nejdl, Leibniz Universität Hannover

2012

“Das Denken für sich allein bewegt nichts, sondern nur das auf einen Zweck gerichtete und praktische Denken.“

Aristoteles, Nikomachische Ethik

Abstract

Controlled content quality especially in terms of indexing is one of the major advantages of using digital libraries in contrast to general Web sources or Web search engines. Therefore, more and more digital libraries offer corpora related to a specialized domain. Beyond simple keyword based searches the resulting information systems often rely on entity centered searches. For being able to offer this kind of search, a high quality document processing is essential.

However, considering today's information flood the mostly manual effort in acquiring new sources and creating suitable (semantic) metadata for content indexing and retrieval is already prohibitive. A recent solution is given by automatic generation of metadata, where mostly statistical techniques like e.g. document classification and entity extraction currently become more widespread. But in this case neglecting quality assurance is even more problematic, because heuristic generation often fails and the resulting low-quality metadata will directly diminish the quality of service that a digital library provides. Thus, the quality assessment of information system's metadata annotations used for subsequent querying of collections has to be enabled. In this thesis we discuss the importance of metadata quality assessment for information systems and the benefits gained from controlled and guaranteed quality.

We start with a short definition of digital libraries and then discus quality concepts in this environment. We continue with the identification where and when quality indicators can and should be measured. For each of the identified areas, we further investigate important aspects of the quality assessment when using semantic techniques. In particular, we present a user study based on a typical semantic technique used for automatic metadata creation, namely taxonomies of author keywords and tag clouds. Based on the evaluation of this experiment, we focused on communalities between the experts' perception and thus draw a first roadmap on how to evaluate semantic techniques by proposing some metrics.

Further, we address problems and possible solutions when using semantic techniques in digital libraries. In particular, we present a user study that on one hand evaluates metrics for quality assessment, and on the other hand evaluates their benefit for the individual user during interaction. To observe the interaction of domain experts, we transferred the abstract metrics' outcome into three different kinds of visualizations and asked experts to evaluate these visualizations. We show that the generated quality information is indeed not only essential for data quality assurance in the creation step of digital libraries, but will also be helpful for designing intuitive interaction interfaces for end-users.

Finally we propose a metadata quality model to determine the overall quality of a metadata set. This model is incorporated into a novel architecture for metadata quality control similar to data warehousing.

Zusammenfassung

Eine kontrollierte Qualität der Metadaten ist einer der wichtigsten Vorteile bei der Verwendung von digitalen Bibliotheken im Vergleich zu Web Suchmaschinen. Auf diesen hochqualitativen Inhalten werden immer mehr fachspezifische Portale durch die digitalen Bibliotheken erzeugt. Die so entstehenden Informationssysteme bieten oftmals neben einer simplen Stichwortsuche auch Objekt zentrierte Suchen an. Um solch eine Objekt-Suche zu ermöglichen, ist eine hochqualitative Verarbeitung der zugrunde liegenden Dokumente notwendig.

Betrachtet man hingegen die heutige Informationsflut, so stellt man fest, dass der Aufwand für eine manuelle Erschließung von neuen Quellen und die Erzeugung von (semantischen) Metadaten für die Indexierung schon heute unerschwinglich ist. Eine aktuelle Lösung für dieses Problem ist die zumeist automatische Erzeugung von (semantischen) Metadaten, durch statistische Methoden, wie die automatische Dokumenten Klassifizierung oder Entitäten Extraktion. Aber bei der Verwendung solcher Methoden ist die Vernachlässigung der Qualität noch problematischer, da eine heuristische Erzeugung oftmals fehlerbehaftet ist. Diese schlechte Qualität der so erzeugten Metadaten wird dabei direkt die Servicequalität einer digitalen Bibliothek mindern. Somit muss eine Qualitätsbewertung der Metadaten garantiert werden. In dieser Arbeit diskutieren wir die Bedeutung von Metadaten Qualität für digitale Bibliotheken und die Chancen die aus kontrollierter und garantierter Qualität gewonnen werden können.

Wir beginnen mit einer kurzen Definition einer digitalen Bibliothek und diskutieren dann Qualitäts-Konzepte in diesem Bereich. Wir fahren mit der Identifikation von Qualitätsindikatoren fort. Für jeden der identifizierten Bereiche untersuchen wir wichtige Aspekte der Qualitätsbeurteilung bei der Verwendung von semantischen Techniken. Insbesondere stellen wir eine Studie vor, die eine typische semantische Technik zur Erstellung von Metadaten begutachtet. Basierend auf den Ergebnissen haben wir Gemeinsamkeiten zwischen den Wahrnehmungen der Experten festgestellt und schlagen darauf aufbauend eine Vorgehensweise für die Evaluierung von semantischen Techniken vor.

Darüber hinaus behandeln wir weitere Probleme und mögliche Lösungen bei der Verwendung von semantischen Techniken in digitalen Bibliotheken. Im Besonderen beschreiben wir eine Studie die zum einen Metriken zur Qualitätsbewertung beurteilt und zum anderen die Vorteile für die Benutzer, die bei der Verwendung dieser Metriken entstehen, untersucht. Wir zeigen, dass die erzeugte Qualitätsinformation in der Tat nicht nur wesentlich zur Sicherstellung der Datenqualität bei der Erzeugung ist, sondern auch hilfreich für die Gestaltung intuitiver Schnittstellen für Endanwender sein kann.

Abschließend schlagen wir ein Qualitätsmodell für Metadaten vor. Dieses Modell wird in einer neuartigen Architektur für die Metadaten Qualitätskontrolle integriert, die ähnlich zum Data Warehousing ist.

Acknowledgements

First, I would like to thank my advisor Prof. Dr. Wolf-Tilo Balke. He introduced me to research, always had time for scientific discussions and gave me the freedom to pursue my research goals. In short, this thesis would not have been possible without his support and guidance.

I also would like to thank Prof. Dr. Wolfgang Nejdl giving me the opportunity to start at L3S Research Center and to be my second referee.

The collaboration and discussion with my colleagues at L3S Research Center and Institute of Information Systems, Technical University of Braunschweig was an essential source of information and has spawned a lot of insights for this thesis. But what is even more important, our joint work was always a pleasure, and we had a lot of fun together. I would like to thank all of my colleagues for their cooperation and openness, especially Benjamin Köhncke, Oliver Koepler, Tereza Iofciu, Kerstin Bischoff, Joachim Selke, Christoph Lofi and last but not least Mohammad Alrifai.

It is immensely helpful to work in a smooth technical environment. Olaf Jansen-Olliges, Marko Brosowski and Dimitar Mitev provided such an environment for L3S, and were always very supportive when I came to them with my minor or major requests.

I will always be grateful for the love and care of my parents. They gave the basis for all my work.

Finally, my wife Sabine and my sons Pit Fide and Jelle Stieg tolerated it with exceeding patience when I couldn't spend enough time with them, and sustained me every day with their love. I thank you for this!

Table of Contents

Chapter 1

Introduction

Searching for literature and information in general is an onerous task not only for domain experts. All of us have to fight the information flood and crave for a perfect personalized knowledge space. Such a knowledge space should have the *reliability of a library* combined with the *benefits of the Web*, i.e. the usage of Web 2.0 and semantic techniques, the scalability and the amount of available information. Nowadays Web search engines like Google, Yahoo, and Bing provide such access, if you define *reliability* in the sense of service uptime. But if we further examine differences between a library and a search engine, we can conclude that a major difference between the two is the quality of service. One could say without doubt that the library is the gold standard in terms of service quality. Libraries are not groundless, even after thousands of years, still used to obtain information. But how could they achieve this quality? Traditionally, libraries have their corpora indexed manually by professional cataloguer. As a result, they produce high quality metadata which can be used for their high quality services.

In fact, such metadata is an essential tool for the focused access to information. Thus, in the field of scientific research portals for literature, such as Web of Science [1], Engineering Village [2], Scopus [3], and GetInfo [4], new technologies like full text indexes and faceted or navigate search has been launched in recent years. On the one hand, these new technologies allow for a fast and accurate search in large data sets. On the other hand, a narrowing of the still very extensive hit list through individual filters, for example, by year of publication, author, or journal title. However, this navigated browsing in the hit list requires the availability of appropriate filter attributes (i.e. metadata) for the document collection. Furthermore, for a satisfactory search quality these attributes must be comprehensive and of high quality.

Certainly with the growth of the Internet as a publication platform and especially with the growing number of open access journals and 'grey literature'/preprint servers, more and more high quality content is made available all the time. A good example for the information growth in highly specialized domains is shown in a press release of the Chemical Abstract Service (CAS) [5]. A total of 50 million chemical substances have been indexed on September 2009 in the curated CAS registry, the worldwide most comprehensive registry of chemical substances. Remarkable is that only 40 million substances have been indexed just nine month before. In contrast, the CAS registry contained 10 million entries in 1990 and around 22 million entries in 2000. Beside the content that is provided by publishers known in a field, also this Web content is increasingly important for digital library users. Hence many digital libraries have extended their collections by harvesting content from the Web. Due to this information flood, the information pro-

viders are continuously facing the challenge of generating high quality metadata in an *automatic way*. Especially enriching the collection of documents available with searchable metadata is necessary for operation. Basically one differentiates between classical, *bibliographic metadata* (such as author, title, year of publication, and publisher) and *semantic metadata* describing the content of a document. Due to the fact that libraries have to make the step from manual to automatic indexing the quality of the generated metadata is questionable.

However, the generation process of semantic metadata is very domain specific. As a result more and more topical portals arise such as ViFaTec[1], ViFaPhys[2], ViFaMath[3], and ViFaChem[4]. But the problem is even more complicated, because the meaning of a metadata field can vary within a domain. Consequently for the metadata generation processes in digital libraries we can state two hypothesis that are fundamental for the work presented in this thesis:

- "The document creation process influences the outcome of semantic techniques for metadata generation and thus the quality of the generated metadata."
- "Depending on the interpretation of the semantic meaning of a metadata field, the quality of the filled-in values may differ a lot for different consumers."

However, unlike the controlled content that arrives from traditional publishers, assessing the quality of harvested content poses severe challenges. Whereas the general quality of each item or Web information source can usually be assessed quite well by the community of users (e.g. using feedback in Web 2.0 interfaces), the business-critical problem of correctly indexing the new content for retrieval with both bibliographic data and content-based index terms remains with the digital library provider. Whereas gathering information from the Web has often been compared to 'trying to drink from a fire hydrant', indexing all this information with controlled quality seems like 'trying to drink from a fire hydrant while assessing the water quality of each sip'. This leads to a trade-off for digital libraries between offering broad and up-to-date document collections and providing high quality metadata for retrieval.

Actually, the *Deutsche Forschungsgemeinschaft* (DFG) emphasized the importance of quality assurance within the information provisioning: "… besides the development of new services also the quality assurance of both the information offers and facilities become more important …" [6]. Also the Bund-Länder Kommission (BLK) mentioned the importance of the underlying quality: "… Not infrequently, the company's success depends on the diligence of the search, so that the com-

[1] http://www.vifatec.de/

[2] http://www.vifaphys.de/

[3] http://vifamath.de/

[4] http://www.chem.de

pleteness of the search plays a greater weight than the precision. This fact entails that the user (e.g. in the chemical and pharmaceutical industry) do not trust fully automated processes and grant them only little acceptance… " [7]. But in practice, during our collaboration with various information providers, we found that the quality is often limited to the lowest common denominator: In most cases only the completeness of a metadata set is audited.

The reason for these circumstances is the lack of research which has been done in the area of quality assurance for metadata quality, semantic technologies and the lack of software which can be used out of the box. All the more, it is questionable, why the well-known digital library frameworks, i.e. BRICKS [8], JeromeDL [9], and DELOS [10], do not consider metadata quality in particular. Only DELOS has the concept of *Quality* specified, but does not support the librarian in analyzing the underlying metadata quality.

1.1. A Motivation: The GetInfo Portal

Consider the example of the GetInfo[5] portal from the German National Library of Science and Technology (TIB). GetInfo is the portal for science and technology and within this portal it is possible to conduct an interdisciplinary search in the stocks of the TIB, the German National Libraries of Medicine and Economics as well as other specialized databases. GetInfo provides access to more than 135 million data sets. The needed metadata is received from many different information sources. Beside commercial publishers, e.g. Springer, Elsevier and Thieme, also Open Access (OA) publishers are considered. With respect to copyright issues, the commercial providers transmit their metadata sets directly to the library, whereas some of the OA publishers open up their repositories via the OAI-PMH protocol or other download possibilities. This content has to be harvested and processed.

Due to the diversity of data providers, every library has to handle a vast amount of different metadata schemas not only while harvesting. Luckily, such schemas do not change a lot over time. Thus, with a number of adequate scripts it is feasible to transform all external schemas to the internal metadata schema. This well-defined internal schema subsequently is an important foundation for the services offered. But even more crucial than a well-defined schema is the quality of metadata filled into the schema [11].

To guarantee the high quality requirements of a digital library while being able to scale with the information growth, currently the document indexing process has to be automated to some degree. The TIB for example uses automated schema transformation tools to transform the received metadata into their own GetInfo schema. Whereas publishers mainly offer *bibliographic metadata*, especially for libraries focusing on specific domains, bibliographic metadata is not enough.

---

[5] https://getinfo.de

Considering the domain of chemistry *semantic metadata* is of high importance. To assist the user's information gathering process in the domain of chemistry it is mandatory to also consider e.g., chemical entities that occur in the respective documents. In practice chemical documents have to be extended by two types of metadata. The first type, the bibliographic metadata like authors, affiliation, publisher and year, is obviously readily available in a library environment. The second and for our purposes more important type is chemical metadata, specified by chemical entities, reactions, concepts and techniques, contained in the original document. This chemical metadata is not readily available and must be extracted, collected and structured. Therefore, the development and application of technologies for automated metadata generation gain in importance. The advantage of using such techniques is twofold: First, document processing becomes less expensive and second, a higher degree of personalization is possible. In particular, the usage of semantic techniques has been proposed to bring a higher rate of automation into the indexing process. Commonly used semantic techniques in the domain of digital libraries are the usage of (bibliographic) ontologies, tagging and classification systems. To offer advanced search experiences the TIB provide a graphical search interface, which has been developed in the course of this thesis, allowing the domain expert to draw the chemical entity of interest. Performing a structure search over the structure database the domain expert will receive a list of chemical entities matching the query. In a next step, the domain expert can choose the entities of interest and a document search over the document repository will be executed. Finally, all related documents are returned and the user has the chance to drill down the result set using bibliographic metadata, like e.g. publication year or author names. Furthermore, also semantic metadata, like, for example, chemical entities and reaction names, are included in the facets [12].

Imagine such chemical metadata is also offered by the respective content providers. The problem is that it is not traceable how this metadata was created. Were the chemical entities extracted automatically from documents, maybe also considering visual representations of structures or data extracted from tables? Were all chemical reactions extracted or only named reactions? These are really important questions because the expected metadata quality highly depends on the extraction process, respectively on the extracted information in case of the reactions.

1.2. Problem statement

Digital Libraries provide a vast amount of digitized information ranging from collections of cultural heritage to specialized topic centered portals. One of the essential differences between digital libraries and unstructured collections, such as the Web, is the focus on information quality. In contrast typical Web search engines base their indexing on text-based measures from information retrieval and structural properties of the collection, e.g. link analysis, whereas digital libraries usually use

indexes (manually) crafted from document metadata. However, this method is no longer applicable due to information overload and has to be switched to automatic processes without neglecting the controlled quality. Only then digital libraries can continue claiming their benefits against search engines, i.e. being the golden standard with regard to service quality.

Today, mostly semantic techniques are used to automatically generate metadata. Such techniques rely on statistical and collaborative methods to assess textual documents. However, due to the nature of statistical and collaborative methods, using such techniques may result in a loss of retrieval quality in comparison to handcrafted indexes. For information providers this potential loss in quality is a serious problem; if users cannot trust in the results, the added value of curated information systems over simple Web searches becomes questionable. Hence, before a semantic technique can be used, information providers have to gauge the impact of the technology's use in the retrieval process. Thereby, the quality of the process itself may not be improved, but the information provider can at least offer a controlled quality and make it visible to the user.

It raises the question, how quality in digital libraries can be expressed. In the context of the evaluation of digital libraries a lot of research has already been done and some frameworks and models have been developed (see Chapter 2). However, all of them have at least on limitation: they assume high quality of the underlying metadata. In spite of the wide agreement on the need to produce high quality metadata, there are fewer consensuses on what *high quality* means and even less on how it should be measured. There are only few metrics and these cannot be applied per se to automatically generated metadata. Thus, in this range there is only one comprehensive work, which permits an evaluation of these metadata. Here, the author state: "… an ideal measurement of metadata quality for fast growing repositories should have two characteristics: to be automatically calculated for each metadata instance inserted in the repository (scalability) and to provide a useful measurement of the quality (meaningfulness)" [13]. But even in this work only bibliographic metadata is considered. With respect to semantic metadata suitable approaches are still missing.

Of course there has been some work done, looking at certain classes of semantic technologies and trying to create quality metrics for them. However, in relation to digital libraries it seems not enough to evaluate the quality of a single technique, but to embed this quality into the library workflow. Since metadata should help users to find, identify, select and obtain resources, the quality of the metadata will be directly proportional to how much it facilitates those tasks. Thus, we can consider quality as the measure of fitness for a task, but this task is very domain dependent and thus quality should not be judged globally.

Summarized, one of the most important problems is the development of a domain independent quality model to express the quality of automatically generated metadata. This model should be easily capable of being integrated in a digital library workflow to support the library in the assessment of the underlying metadata.

To investigate this problem we will consider the entire life cycle of (semantic) metadata starting from the preprocessing and ending with the document retrieval. For each step within this workflow, we highlight the main pitfalls and provide some techniques within but not limited to the domain of chemistry. We will present among others

— a generic evaluation method for quality judgments of semantic techniques
— a generic metadata quality model
— a proposed architecture for a quality driven digital library

1.3. Thesis Structure

In Chapter 2 we start with detailed reviews of related work in the context of the quality of *traditional* digital libraries. We than extend the quality metrics to semantic digital libraries and discuss related work already done in this area (see 2.1). Extending the quality concept to metadata quality, we discuss related work done in this area in 2.2. This chapter is concluded in section 2.3.

In Chapter 3 we first motivate our use case scenario, i.e. chemical digital libraries, in section 3.1. Afterwards we describe the current state of the art of the information seeking process in the chemical domain.

Chapter 4 introduces an information life cycle which is used to identify where and when quality indicators can be measured (see 4.1). Based on these findings, we discuss quality assessment problems and solutions during a typical workflow in a digital library: document preprocessing (see 4.2), semantic metadata enrichment (see 4.3), document indexing (see 4.4) and retrieval (see 4.5). This chapter is concluded by introducing a life cycle model for the overall digital library quality (see 4.6). Based on this model and the findings in chapter 4, we come up with a quality model (see 5.2). This model is than incorporated in a novel architecture / workflow for semantic digital libraries (see 5.3).

Chapter 6 concludes the thesis with an enumeration of the contributions, while also discussing possible future research directions and open challenges associated with these topics.

1.4. Publication Overview

The ideas and algorithms presented in this thesis have been published and thus peer reviewed at various conferences. We describe contributions included in:

- S Tönnies, B. Köhncke, W.-T. Balke, "Meta-Line: Lineage Information for Improved Metadata Quality", *12th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Washington DC, USA: 2012.
- S. Tönnies, B. Köhncke, and W.-T. Balke, "Taking Chemistry to the Task - Personalized Queries for Chemical Digital Libraries," *11th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Ottawa, Canada: 2011.
- S. Tönnies and W.-T. Balke, "Quality Assessment in Digital Libraries - Challenges and Chances," *Proceedings of the 22. GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), Bad Helmstedt, May 25-28, 2010*, 2010.
- S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Exposing the Hidden Web for Chemical Digital Libraries," *10th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Surfers Paradise, Gold Coast, Australia: 2010.
- S. Tönnies and W.-T. Balke, "Uncovering Hidden Qualities - Benefits of Quality Measures for Automatically Generated Metadata," *14th European Conference on Digital Libraries, ECDL 2010, Glasgow, Scotland, September 6 - 10, 2010*, Berlin, Heidelberg: Springer Berlin / Heidelberg, 2010.
- S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Building Chemical Information Systems - the ViFaChem II Project," Datenbanksysteme in Business, Technologie und Web (BTW 2009), 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), GI, 2009.
- S. Tönnies and W.-T. Balke, "Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation," *13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009*, Berlin, Heidelberg: Springer Berlin / Heidelberg, 2009, pp. 168-179. *Best Paper Nominee*

During the ViFaChem II project several other aspects, which are not part of this thesis, have been published in the area of Chemical Digital Libraries.

- B. Köhncke, S. Tönnies, W.-T. Balke, "Catching the Drift – Indexing Implicit Knowledge in Chemical Digital Libraries", *International Conference on Theory and Practice of Digital Libraries (TPDL)*, Paphos, Cyprus, 2012
- O. Koepler, W.-T. Balke, B. Köhncke, and S. Tönnies, "Personalised information spaces for chemical digital libraries", *Chemistry Central Journal*, Vol. 3, 2009, p. P41.
- S. Tönnies and W.-T. Balke, "User-centered Content Provisioning over Large Collections of eBooks," Proceedings of the 2009 2nd ACM Workshop on Research Advances in Large Digital Book Repositories, BooksOnline 2009, Corfu, Greece, October 2, 2009, 2009.

- O. Koepler, W.-T. Balke, B. Köhncke, and S. Tönnies, "Perzonalised Information Spaces: Improved Access to Chemical Digital Libraries," *5th German Conference on Chemoinformatics, Goslar, Germany, November 2009*, 2009.

During the early stages of the Ph.D. studies I have also published a number of papers investigating personalization in the area of multimedia Web services and Web modeling.

- S. Tönnies, B. Köhncke, P. Hennig, I. Brunkhorst, and W.-T. Balke, "A Service Oriented Architecture for Personalized Universal Media Access," *Future Internet*, vol. 3, Apr. 2011, pp. 87-116.
- S. Tönnies, B. Köhncke, P. Hennig, and W.-T. Balke, "A Service Oriented Architecture for Personalized Rich Media Delivery," *2009 IEEE International Conference on Services Computing*, IEEE, 2009, pp. 340-347.
- I. Brunkhorst, S. Tönnies, and W.-T. Balke, "Multimedia Content Provisioning Using Service Oriented Architectures," *2008 IEEE International Conference on Web Services*, IEEE, 2008, pp. 262-269.
- A. Bozzon, T. Iofciu, W. Nejdl, and S. Tönnies, "Integrating Databases, Search Engines and Web Applications: A Model-Driven Approach," *7th International Conference, ICWE 2007 Como, Italy, July 16-20*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 210 - 225.
- A. Bozzon, T. Iofciu, W. Nejdl, A.V. Taddeo, and S. Tönnies, "Role based Access Control for the interaction with Search Engines," *1st International Workshop on Collaborative Open Environments for Project-Centered Learning, September 2007, Crete, Greece*, 2007.

Before starting my Ph.D. I published some work in the area of vehicle navigation systems and extreme programming.

- S. Tönnies, „Zielführung in der Fahrzeug-Navigation mittels Mixed Reality." Vdm Verlag Dr. Müller, Saarbrücken (2008).
- S. Tönnies, „Neue Techniken in der Fahrzeugnavigation: Mixed Reality als Schlüsseltechnologie.", In: GIS-BUSINESS Geoinformationstechnologie für die Praxis. 08, 2007.
- C. Brenner , V. Paelker , S. Tönnies, „Zielführung in der Fahrzeug-Navigation mittels Mixed Reality", In: Stefan Müller, Gabriel Zachmann (Hg.): Virtuelle und Erweiterte Realität. Proceedings zum 3. Workshop der GI-Fachgruppe VR/AR, Shaker Verlag, 2006.
- Gößner, J., Tönnies, S., Steimann, F. Projectory - ein Tool zur Unterstützung des Einsatzes von XP Techniken im universitären Programmierpraktikum. In: Kerres, M., Witt, C.D., Kalz, M., and Stratmann, J. (eds.) Didaktik der Notebook-Universität. pp. 203 - 219. Waxmann; Auflage: 1., Aufl., Münster (2004).

Chapter 2

Quality in Digital Libraries

In this chapter we will discuss the term *quality* in relation to digital libraries. But before discussing the concept of quality, we have to answer the question: What is a digital library? The formal 5S framework from [14] tries to answer at least partially this question. According to the authors a digital library can be defined by the following 5S:

1. *Streams*: Streams are sequences of elements of an arbitrary type (e.g., bits, characters, images, etc.). With this in mind, they can model static (e.g. text) and dynamic (e.g. video) content.
2. *Structures*: A structure specifies the way in which parts of a whole are arranged or organized. In digital libraries, structures can represent hypertexts, taxonomies, system connections, user relationships, and containment.
3. *Spaces*: A space is a set of objects together with operations on those objects that obey certain constraints. The combination of operations on objects with the set of objects is what distinguishes spaces from streams and structures.
4. *Scenarios*: One important type of a scenario is a story that describes possible ways to use a system to accomplish some function that a user desires.
5. *Societies*: A society is a set of entities and the relationships between them. The entities include humans as well as hardware and software components, which either use or support digital library services. Societal relationships make connections between and among the entities and activities.

Further the authors build a taxonomy of digital library concepts derived from the literature. For the most important concepts, they developed formal definitions (in total 24) and mapped them (see Fig. 1).

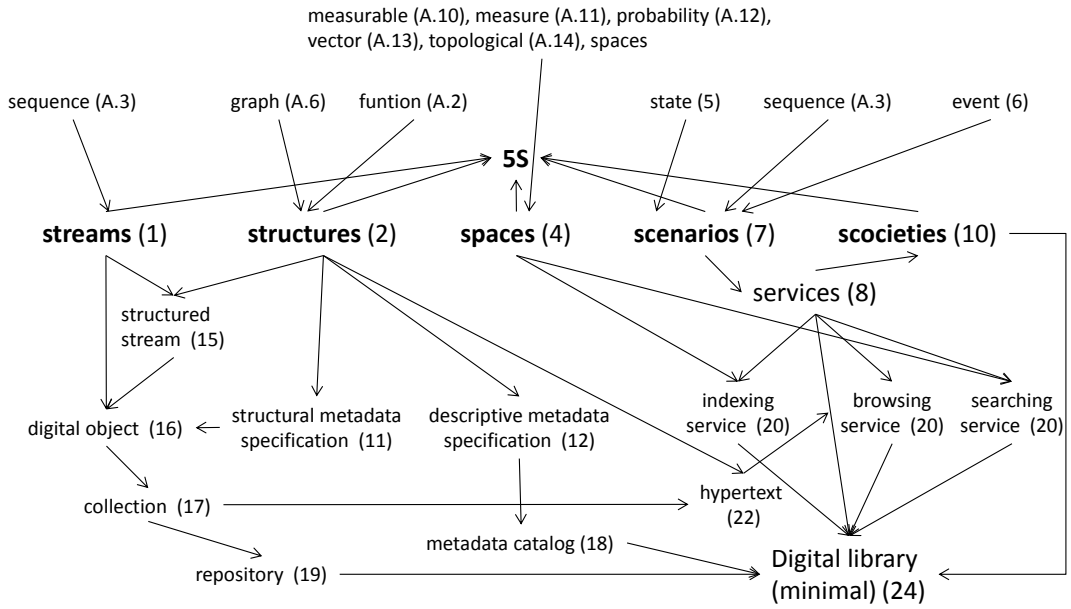


Fig. 1. 5s Map of Formal Definitions [14]

2.1. Evaluating Quality in (Semantic) Digital Libraries

What defines high quality digital library? Already before the introduction of the 5S framework, Saracevic was one of the first authors to consider this problem [15]. He argues that any evaluation basically raises issues such as the *criteria*, the *measures*, the *context* and the *methodology*. However, his analysis shows that there is no agreement regarding the exact elements of these issues for digital library evaluation. Trying to fill some gaps in this area, Fuhr et al. developed a new description scheme using four major dimensions: *collection*, *technology*, *users* and *uses* [16]. Based on this dimensions, a questionnaire was developed and the need for an appropriate test collection was stated, similar to the TREC and CLEF initiatives. Still, both approaches are defined in very general terms maybe because of the lack of the formal definition of a digital library.

Much more precise, because deeply grounded in the formal 5S framework [14], is the work by Gonçalves et al. [17]. The authors identified several digital library key concepts (out of the 24 defined in the 5S framework) to be considered a digital library through a comprehensive review of related literature. In a next step, they made a comprehensive literature review on quality issues in information systems and derived several dimensions of quality. Finally, for each of these dimensions, the variables to measure, together with the respective S were identified. Table 1 connects their proposed dimensions with some "S"-related concepts involved in their definition.

Table 1. Dimensions of quality and S involved in their definitions (taken from [17])

Digital Library Concept	Dimension of Quality	Some "S" concepts involved
Digital Object	Accessibility	Societies (actor), Structures (metadata specification), Streams + Structures (structured streams)
	Pertinence	Societies (actor), Scenarios (task)
	Preservability	Streams, Structures (structural metadata), Scenarios (process (e.g., migration))
	Relevance	Streams + Structures (structured streams), Structures (query), Spaces (Metric, Probabilistic, Vector)
	Similarity	Same as in relevance, Structures (citation/link patterns)
	Significance	Structures (citation/link patterns)
	Timeliness	Streams (time), Structures (citation/link patterns)
Metadata specification	Accuracy	Structure (properties, values)
	Completeness	Structure (properties, schema)
	Conformance	Structure (properties, schema)
Collection	Completeness	Structure (collection)
Catalog	Completeness	Structure (collection)
	Consistency	Structure (collection)
Repository	Completeness	Structure (collection)
	Consistency	Structure (catalog, collection)
Services	Composability	See Extensibility, reusability
	Efficiency	Streams (time), Spaces (operations, contraints)
	Effectiveness	See Pertinence, Relevance
	Extensibility	Societies + Scenarios (extends, inherits_from, redefines
	Reusability	Societies + Scenarios (includes, reuses)
	Reliability	Societies + Scenarios (uses, executes, invokes)

The first comprehensive study on digital libraries evaluation frameworks is presented in [18]. The attractiveness of the collections and the technology's ease of use are identified as key factors in assessing the quality of a digital library. Moreover, the importance of the user satisfaction is emphasized. The model presented is

the *interaction triptych model* which defines three components of the digital library: the system, the content, and the user. In addition three axes of evaluation were provided: usability of user interaction with the system, usefulness of the content for the user, performance of managing the content by the system. Recent research is trying to adopt Web metrics, originally developed for evaluating e-commerce applications, for evaluating digital libraries [19]: preliminary results discuss, e.g., the usage of session length for evaluating the customers' satisfactions with the portal.

With upcoming semantic digital libraries like JeromeDL [9] the question of quality has to be extended: what defines a high quality *semantic* digital library? Kruk et al. do not really answer this question when evaluating JeromeDL against a standard digital library measuring several traditional aspects like precision / recall and the user satisfaction [20]. The conducted user studies imply that the individual user's satisfaction seems to be higher when using semantic technologies. However, it has to be pointed out that the results cannot be generalized, since semantic techniques are just as good as the underlying metadata.

Summarizing this subsection, we can state, that there are some approaches to determine the quality of a digital library. Some of them are high level approaches and thus manual inspection is need for the quality assessment. The others are very formal and thus may be used by computer systems to determine the quality. But all of them have in common, that they assume good quality of the underlying metadata which have never been a problem because this metadata was traditionally handcrafted. However, we have illustrated in the introduction that this approach is not feasible with the ever growing amount of information. Therefore, we must first analyze how we can determine the quality of the underlying metadata in order to finally determine the quality of a digital library.

2.2. (Meta-)data quality evaluation

In general the metadata quality is essential for the operation [21], [22] and interoperability [23], [24] of digital libraries. The services in a digital library can be heavily compromised by low-quality metadata. The metadata should contain enough information, so that the user can obtain a first impression of the described resource without directly accessing the resource itself. For example, considering a digital library aggregating documents from various open access journals, the correctness of the URL where the resource can be accessed is essential for the usefulness of the digital library. This is even more important considering pay per view offers in digital libraries. Thus, the importance of metadata has always been an integral part of resource cataloging [25]. Nonetheless, most implementations still rely on the assumption that metadata was created by an expert in the specific field and should have an acceptable degree of quality (see also [13]).

Table 2. Overview of different metadata quality evaluation studies

Study	Approach	Instances	Main focus of evaluation
Greenberg et al. [26]	Manual	11	Quality of non-expert metadata
Moen et al. [27]	Manual	80	Overall quality of instances
Shreeves et al. [28]	Manual	140	Overall quality of instances
Stvilia et al. [29]	Manual	150	Identify quality problems
Wilson [30]	Manual	100	Quality of non-expert metadata
Bui and Park [31]	Statistical	1.040.034	Completeness of instances
Hughes [32]	Statistical	27.000	Completeness of instances
Najjar et al. [33]	Statistical	3.700	Usage of the metadata standard

Today, with growing repositories, quality issues become more apparent, leading to the adaptation of techniques developed to review the quality of physical library instances. These approaches can be split into two groups: manual quality evaluation and simple statistical quality evaluation (see Table 2). The first set of approaches *manually reviews* a statistically significant sample of the metadata instances against a predefined set of quality attributes. Manual evaluations are done by humans and averaged for an estimation of the metadata quality in the repository. For example, [26] reports on a study that examined the ability of resource authors to create acceptable quality metadata in an organizational setting using manual evaluation by experts. So far these methods are the most meaningful way to measure metadata quality in a digital library. Indeed they have three major disadvantages whereby they have little practical impact and are mainly research activities:

1. The manual quality assessment is only valid at sample time. If a notable amount of resources is added to the digital corpus, the assessment could be invalid and the estimation has to be redone.
2. The quality of individual metadata records can only be obtained for records contained in the sample. For the whole set, only the average quality can be obtained.
3. This kind of quality assessment is still costly because humans have to review a huge number of objects that is always increasing. Thus, these methods have only limited advantage in comparison with the manual creation of metadata records.

The second set collects *statistical information* about all the metadata instances in the repository to determine an estimation of their quality. In more detail, they conducted a statistical analysis on a sample of metadata records from various re-

positories and evaluate the usage of the standard. They designate the most frequently used fields and values attributed to these fields. While not directly associated with quality, the statistical indices produced provide an insight of the efficiency of the repository examined without the cost involved in manual quality review. Unfortunately they do not provide a similar level of "meaningfulness" as a human generated assessment.

A more systematic and organized view of metadata quality with less subjectivity is achieved with the introduction of generic frameworks for the evaluation of quality. In general, these frameworks introduce parameters that indicate whether information should be considered of high quality. The frameworks differ a lot in their scope and goals. They have been inspired by the Total Quality Management paradigm [34], text document analysis [35], and research on library catalogs [36]. Because they are directly related to this thesis we will focus on the last group. In [27] a procedural framework for evaluating metadata records using 23 evaluation criteria is introduced. The framework discussed in [37] is based on concepts and ideas of the more generic field of information quality. It identifies 32 information quality parameters classified into 3 dimensions of information quality: intrinsic, relational/contextual and reputational. Bruce and Hillman [38] condense many of the parameters in order to improve their applicability. They elaborate 7 characteristics of metadata quality: *completeness*, *accuracy*, *provenance*, *conformance to expectations*, *logical consistency and coherence*, *timeless,* and *accessibility*. Finally, the authors of [28] put both frameworks into relation (see Table 3).

However, some proposals are available for determining metadata quality automatically. For example, the authors of [13] use the theoretical background of the Bruce & Hillmann framework and attempt to operationalize the measurement of quality in a set of automatically calculated metrics for the 7 parameters. Similar efforts to provide metrics for metadata quality measures can be found in [32].

Today, the Bruce & Hillmann framework is effectively the most used in the domain of digital libraries. This is not just the fact because the seven characteristics are easy to understand by humans but also that it keeps all dimensions of quality proposed in other frameworks. However, all these frameworks have a static metadata instance in mind, thus they are less appropriate for digital libraries than for traditional libraries. Since we will focus on digital libraries and highly dynamic environments in the sense of metadata updates we have to tackle the quality assessment of *dynamic metadata*. This kind of metadata can change whenever the metadata resource is used or accessed. Given that, to the best of our knowledge, there is no framework available to describe the quality of dynamic metadata. Even worse, when working with semantic techniques, the quality of their outcome and thus the generated metadata is questionable due to their statistical nature.

Table 3. Mapping between the Bruce & Hillman [38] and the Stvilia et al. framework [37] (reprinted from [28])

Stivila et al.		Bruce & Hillmann
Dimension	**Parameters**	**Characteristics**
Intrinsic	Accuracy / Validity	Accuracy
	Cohesiveness	Logical consistency and coherence
	Complexity	Accessibility
	Currency	Timeless
	Informativeness	Completeness
	Naturalness	Accessibility
	Precision	Conformance to expectations
	Semantic consistency	Logical consistency and coherence
	Structural consistency	
Relational	Accuracy	Accuracy & Conformance to expectations
	Completeness	Completeness & Conformance to expectations
	Complexity	Accessibility & Conformance to expectations
	Informativeness	Conformance to expectations
	Latency / Speed	Accessibility
	Naturalness	Accessibility & Conformance to expectations
	Precision	Conformance to expectations
	Relevance	
	Security	Provenance
	Verifiability	Provenance & Conformance to expectations
	Volatility	Timeless
Reputational	Authority	Provenance

Use Case Study: Evaluating the Semantic GrowBag

Let us consider a typical way of accessing digital collections. Since metadata in the form of descriptive terms is often used to describe and summarize documents, it is often also offered for navigational access. Such terms can either be provided by the documents' authors, but can also be derived from controlled vocabularies, e.g. by the publisher. The collections allow users to browse documents based on the keywords organized by some categorization system or thesaurus, i.e. searches can be broadened by choosing more general terms or focused by using more specific terms. However, creating and maintaining the underlying categorization systems is mostly done manually with very high efforts and they are often only available for specific domains.

To limit these efforts recently semantic techniques to automatically created categorization systems in the form of taxonomies have been proposed. Examples are statistical evaluation of term co-occurrences [39], language models [40], or syntactical contexts [41]. Although such techniques allow the automatic creation of taxonomies, the suitability of the resulting classification system for actually searching documents is problematic. How can the quality of such generated taxonomies be assessed? While for Web search rephrasing queries in different terms are acceptable, users of digital libraries expect clear and efficient navigation paths. Hence, the measuring of classification systems' quality becomes an important part in the adoption of semantic technologies.

The actual measurement widely varies in semantic technology research ranging from manual inspection (of random partitions) of the taxonomy to comparison of the entire taxonomy with some kind of 'gold standard'. For instance, in the area of (bio-) medical collections the MeSH taxonomy [42] provides an often used benchmark: when putting an implementation to the test it is run over a focused collection like e.g. the Medline corpus [42] and the resulting taxonomy is compared to the corresponding MeSH entries and their respective relationships. For example, in [43] a technique called Semantic GrowBag (based on term co-occurrences, for details see [44]) is used to compute more than 2000 individual taxonomies over Medline documents. It is interesting to notice that for deriving sensible topical taxonomies a minimum of about 100,000 documents was necessary, since statistical methods only provide meaningful results using a sufficiently large sample. For evaluation subsequently the average percentage of accordance or discrepancy with respect to MeSH is presented. Still, it is not clear what these percentages mean in terms of the libraries usability when the respective taxonomies are used as classification system for navigational access.

Additionally, evaluation metrics used for semantic technologies and their generated metadata are bound to a specific technology. Therefore a lot of different metrics have been proposed. Particularly in the domain of collaborative tagging systems, some work investigating tag quality has been performed. According to [45] the distributions of different tags for each individual document tend to stabilize

over time, i.e. more and more users add meaningful tags whereas irrelevant tags are not amplified. This result is confirmed in [46] and the authors show in addition, that tags follow a power law distribution. Taking these properties into account, it seems likely that user tags gathered in an unsupervised fashion can, indeed be a reliable source of information [47], [48].

For searching and metadata creation within tagging systems, [49] proposes the exploitation of co-occurrence of users, resources, and tags. This is done using a graph model to represent the folksonomy. In [47] tag data is explored for the purpose of Web search through the use of two tag based algorithms: one exploiting similarity between tag data and search queries, and the other one utilizing tagging frequencies to determine the quality of Web pages. Chan examined a huge number of query terms posed to Powerhouse and concludes that the combined usage of folksonomies with taxonomies increases the recall of the information seeking process [48]. In contrast [50] found out that the use of only document terms yielded slightly better F-measure than using terms and tags together. The authors' results suggest that not all tags are useful descriptors for resource sharing. This leads to the question which kind of tags have a high quality: Bischoff et al. [51] showed that it is worthwhile having a common tag classification scheme for different collections – allowing tags to be compared tags used in different tagging environments. The experiments show that more than 50% of all existing tags bring new information to the resources they annotate and that a large amount of tags are accurate and reliable. A general algorithm for measuring the quality of tags is proposed in [52]. The authors decoupled the relationship between users and tag-resource pairs modeling the tag-resource pairs as nodes and co-user relationship as edges of a graph. This structure allows every two tag-resource pairs used by the same user to have different quality. The algorithm then propagates quality scores iteratively through the graph after being initialized with a set of seed nodes. Still, no group investigates the quality of the resulting tag sets per resource. In particular there is no work comparing the added value in terms of metadata for manually annotated document collections (like curated digital libraries) with collaboratively annotated collections. However, this quality gap needs to be measured, if the metadata quality of automatically annotated digital collections like proposed in [53] is to be determined.

In categorization systems, especially ontologies, much work has been done, and several metrics for assessing the quality of an ontology have been proposed, e.g. QOOD [54], OntoMetric [55], and OntoQA [56]. However, all these metrics remain purely on the structural level of the ontology, which is according to [57], not sufficient. In particular, the semantic quality, in terms of correctness, has to be addressed and the authors propose the development of semantically aware ontology metrics. As a first step the authors define the normalization of ontologies and introduce the term of stable metrics. The measurement of the semantic of an ontology becomes essential considering automatically generated ontologies.

These findings reinforce that defining a "one size fits all" quality metric is not a solution and that the *proper metrics for semantic techniques are highly domain specific*. Rather, assessing data quality is an on-going effort that requires awareness of the fundamental principles underlying development of subjective and objective data quality metrics. Hence, in [58] the authors presented subjective and objective assessments of data quality metrics, as well as three functional forms that can help in developing data quality metrics. Finally, the author of a survey about metadata quality in digital repositories [59] conclude that still accuracy, completeness, and consistency are the most commonly used criteria in measuring metadata quality. He states, that there is a pressing need for the building of a common data model that is interoperable across libraries and that the development of a framework for measuring metadata quality and mechanisms for improving quality are also critical areas for further studies.

2.3. Conclusion

The quality of a digital library has always been very strongly linked to the satisfaction of the users. This cutback of the quality concept has been valid since manual annotation of metadata was implied. However, this is no longer appropriate as more and more metadata is generated automatically. Hence, the quality concept has been extended also considering quality of automated generated metadata. One of the most appropriate work has been done by Ochoa et al. [13]. Based on the parameters identified by Hillman et al. [38] they provide 7 metrics (completeness, accuracy, conformance to expectation, consistency, accessibility, timeliness, and provenance) which can be computed automatically to judge the quality of metadata instances. These general quality metrics should be applicable for all digital collections. This may be true for bibliographic metadata, but it remains questionable whether it is also true for semantic metadata, which is very domain specific by nature. Generally it can be observed that there are more and more metrics available to assess the quality of specific semantic techniques. Thereby, in most cases, those metrics cannot be applied to different semantic techniques. Accessorily, these metrics seem to be very domain dependent. Thus, it cannot give a one-size-fits-all solution, as it has been postulated in previous work.

It seems obvious that we need a higher-level quality model, in which all the metrics can be chosen based on the underlying technology and metadata standard and still allow a statement about the quality of the resulting metadata. It is important that the quality can always be measured with respect to a task, as we have already shown that the quality of metadata is always considered in relation to their task needs.

Chapter 3

Chemical Digital Libraries as Use Case

Different scientific disciplines have their own demands and the respective community has different workflows and expectations when it comes to searching for literature. Hence libraries have branched out into topically centered virtual libraries for several disciplines closely focusing on the needs of each individual science. That is one reason, why we stated that quality metrics for semantic techniques are domain dependent. For an illustration of this problem, we have chosen chemistry because keeping the risks and extremely high costs for research and development in chemical and pharmaceutical industry in mind it is obvious that this domain highly relies on quality information: missing one important publication can compromise the whole work of a research project.

3.1. Motivating the Use Case: A Typical Chemical Workflow

The following scenario showcases the daily tasks of a researcher in the chemical domain. Let us focus on drug design, which usually includes aspects of isolation of natural products, their structure elucidation, their pharmacological activities, their synthesis and finally the search for new derivatives with enhanced properties for a certain mode of action. Assume our scientist is interested in anti-cancer drugs, particularly the class of taxanes see e.g. [60]. He or she may start by looking for information about Paclitaxel (see Fig. 2) and related drugs. Paclitaxel (often referred to under the brand names 'Taxol' or 'Abraxane') is a terpenoid isolated from the bark of yews, with a very high activity against several tumor cell lines and thus is of high interest for pharmaceutical research since its first isolation and structure elucidation in 1971. Naturally our researcher is especially interested in the mode of action of Paclitaxel and maybe other compounds with similar properties. Furthermore he is looking for experimental procedures for the synthesis of Paclitaxel-like structures or precursors.

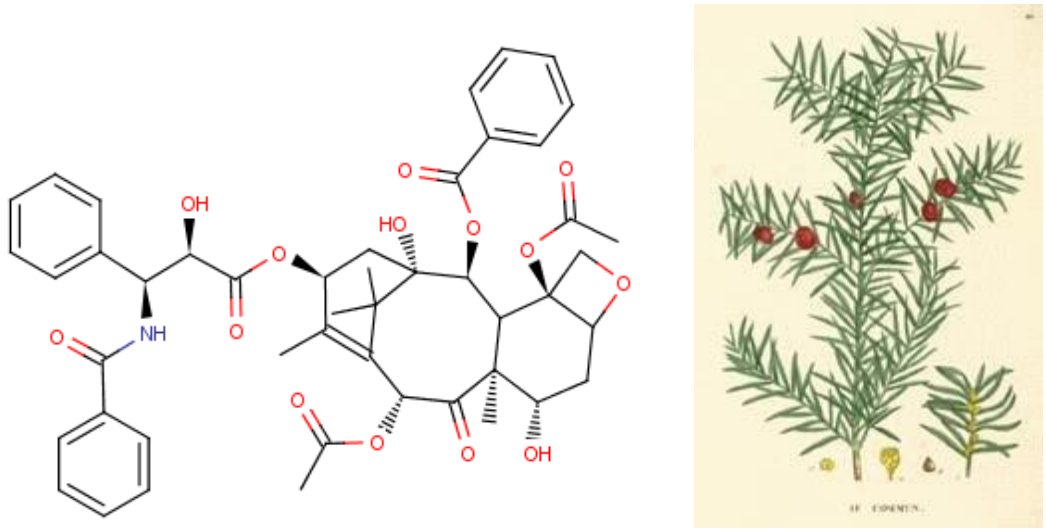


Fig. 2. Structure of Paclitaxel (left) isolated from the yew tree (right)
(botanical image from: M.Grieve. 'A Modern Herbal', Harcourt, Brace & Co, 1931)

The common information retrieval process of a text-based search as known from other domains, will fail for this scenario for many reasons: the questions of the researcher involve information about chemical entities, concepts and facts. But as stated above queries involving chemical structure information either in form of substructures or similar structures can hardly be expressed in form of keywords. Of course, searching for the chemical entity name 'Paclitaxel' may return some results, but as a non-proprietary name it may not be used broadly in scientific research papers. One can try the IUPAC name of Paclitaxel as generated by a large ruleset published by the IUPAC. But especially for complex molecules there are several ways how to interpret the IUPAC guidelines for nomenclature, so one still does not have a unique identifier for the molecule.

The only unambiguous representation of the entity Paclitaxel is its structural representation. Over the years several line annotations have been developed, which allow the conversion of graphical structure information into strings. These algorithms provide compact strings representing chemical entities, but are not easy to read and interpret by humans. Especially for complex structures line annotations become difficult to handle. The following lines show both the SMILES and the InChI code for Paclitaxel.

InChI: InChI=1/C47H51NO14/c1-25-31(60-43(56)36(52)35(28-16-10-7-11-17-28)48-41(54)29-18-12-8-13-19-29)23-47(57)40(61-42(55)30-20-14-9-15-21-30)38-45(6,32(51)22-33-46(38,24-58-33)62-27(3)50)39(53)37(59-26(2)49)34(25)44(47,4)5/h7-21,31-33,35-38,40,51-52,57H,22-24H2,1-6H3,(H,48,54)/t31-,32-,33+,35-,36+,37+,38-,40-,45+,46-,47+/m0/s1/f/h48H

SMILES: CC1=C2C(C(=O)C3(C(CC4C(C3C(C(C2(C)C)(CC1OC(=O)C(C(C5=CC=CC=C5)NC(=O)C6=CC=CC=C6)O)O)OC(=O)C7=CC=CC=C7)(CO4)OC(=O)C)O)C)OC(=O)C

One can easily see that these identifiers are difficult to handle in a text-based search and are no alternative to a semantic rich drawn chemical structure. Using a retrieval system with a graphical user interface our researcher can easily draw the molecular structure of Paclitaxel based on the rulesets of chemistry. Moreover, a structure based search is essential when it comes to the search for similar structures or structures which contain residues of a given lead structure.

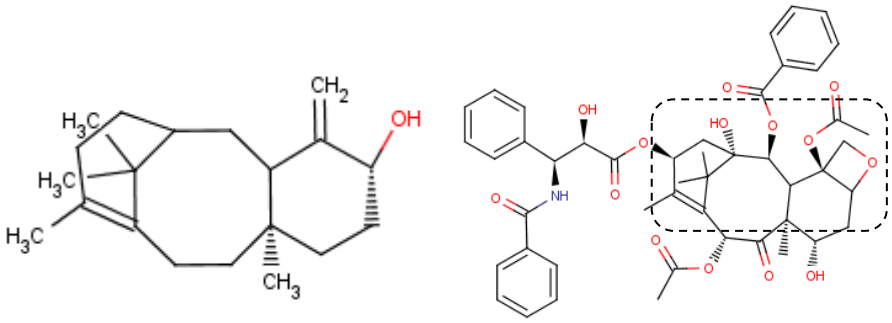


Fig. 3. Chemical structure of Taxadien-5-α-ol (left)
and structurally similar group in the Paclitaxel molecule (right)

In our case the researcher may find out, that taxadien-5-α-ol (cf. Fig. 3) is a central precursor in the biological synthesis of Paclitaxel. A good retrieval system now has to offer search capabilities for known molecules containing this particular skeleton with the 8-membered ring. At the latest at this point in time a text-based keyword search has become entirely useless. Such chemical structure related information retrieval can only be handled with a substructure search process. The same paradigm is valid for queries concerning the synthesis of the target molecule or precursors containing essential fragments or functional groups. All these operations are based on the information how atoms of the molecule are connected. However, the depth of structural information may vary: whereas the simplest representation contains only information about the composition of a molecule, a topological representation will also contain information about the connectivity of a molecule, showing which atoms are connected by which bond type. Moreover, the topographical representation can comprise spatial arrangements of atoms and bonds, showing the stereochemistry and conformation of a molecule, too. Therefore structure databases used for chemical searches generally contain information on the level of topological representations. In contrast to the basic molecular formula, two-dimensional representations are the natural language of chemistry, a universal code understood by all chemists worldwide.

Thinking of integrating taxonomical information for the chemical domain (for query expansion or recall-oriented queries), a retrieval system should also offer a navigational access over the related substance classes of the chemical entity. Thus, helpful information about structural superclasses and related substance classes is provided. Our researcher may know that Paclitaxel belongs to the toxoids (see Fig. 4), which are in turn diterpens with a taxan-like skeleton. The diterpens on the other hand are built up from 4 isoprene units always having 20 carbon atoms and can be divided into acyclic, monocyclic, tri- and tetracyclic diterpens, Paclitaxel is a member of the later tetracyclic diterpens. The most prominent member of these tetracyclic diterpens is Phorbol, which interestingly is a strong carcinogen.

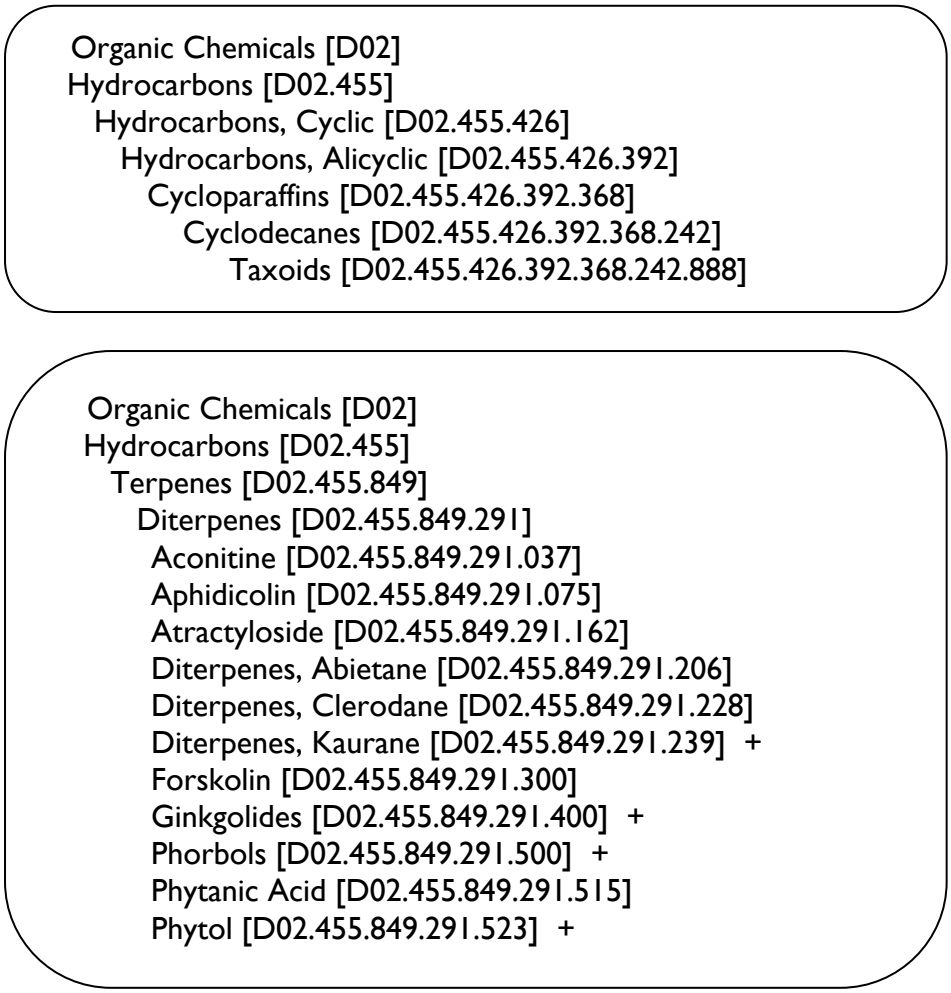


Fig. 4. Taxonomical information about Paclitaxel from MeSH

Moreover, since our researcher is interested in drug design also information about the medical use of Paclitaxel will be used to retrieve documents and structure relevant information. For example in the MeSH ontology of the US National Center for Biotechnology Information (NCBI) the pharmaceutical action of

Paclitaxel is classified as 'antineoplastic agent, phytogenic'. Furthermore it is approved by the US Food and Drug Administration (FDA) to treat ovarian cancer and breast cancer.

3.2. Searching for Chemical Knowledge – State of the Art

Concluding the previous section, we can state that for chemical literature it is not sufficient just to provide keyword-based access. Chemical information deals with information about molecules, their properties, their biological and pharmaceutical activities, their industrial use, and their reactions. To a large degree chemical information is communicated by structures or (taxonomies of) substance classes instead of verbal descriptions and practitioners can very efficiently discriminate between substances based on their visual representations or super classes. For a high quality information retrieval it is therefore important to cover the information provided about chemical entities based on actual chemical workflows. In order to do that strong interdisciplinary work is mandatory. In particular many of the current approaches for handling chemical information in computer science are entirely contrary to chemical workflows and thus have a rather theoretical impact on the field.

Chemical search engines specialized on *chemical formulas* are one example of searching for chemicals in text documents, which disregards the specific demands of the chemistry domain. While organizing documents by indexing formulas of chemical substances occurring in the document seems reasonable from a computer science point of view (for a recent example see e.g., [61], [62]), from a chemical point of view chemical formulas are highly ambiguous and definitely not the right choice for representing chemical entities for search. A chemical formula like $C_6H_{12}O_6$ of a molecule provides only information about the elements contained in this molecule and their respective number of atoms. There is a variety of possibilities to instantiate the actual molecule: for instance $C_6H_{12}O_6$ primarily represents the sugars glucose, fructose and mannose. But there are 21 more sugars, all with the identical chemical formula. Actually, based on bonding rules there are 267.258 theoretical molecules matching $C_6H_{12}O_6$, not including stereoisomers[6]. Even worse, if $C_6H_{12}O_6$ is just given as a fragment for a substructure search, the possibilities of identifying matching substances are overwhelming.

Using *entity names* for substances is also difficult. Although the standardized IUPAC name, (2S,3R,4S,5R,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol is rarely used for D-glucose compared to the more prominent synonyms like dextrose, corn sugar, or grape sugar. On the other hand a *structural formula* defines a graphical representation of a molecular structure showing atoms, bonds and their spatial arrangement, describing a unique chemical entity. Although structural formula are

[6] Based on MOLGEN, http://www.molgen.de/

presented as graphical information there are textual descriptors like InChI[7], SMILES[8] or CML [63] providing structural information as strings. Chemical information is therefore often related in the form of InChI or SMILES codes, the (commercial) CAS registry number, fingerprints or a detailed structural description. For example D-Glucose has the CAS registry number 492-62-6, and the InChI string: 1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1.

We can thus see that only interdisciplinary work will lead to a high quality information provisioning platform that is promising to be accepted by a wide range of practitioners in the field. In fact, already during the nineteenth century, inspired by the work of Jacob H. van't Hoff and August Kekulé, drawings of chemical structures became the common way of communicating chemical information about substances and their reactions. Today, we speak of chemical structure representations as the 'language of chemists' [64]. The chemical structure is a simple to understand, yet most precise way to uniquely describe a chemical entity, leaving the ambiguity of systematic, IUPAC, trivial or brand names behind. Graphical representations of chemical entities are therefore commonly used as query terms in searching for chemical information. However, although easily recognized by the human eye, graphical representations of chemical entities still cannot be easily transferred into the digital world once published in a document.

Over the last years, several projects focused on developing a chemical optical recognition for the reconstruction of chemical structure information from digitized documents. However, recognition rates always have proven to be insufficient in a production environment [65–68]. That's why the most comprehensive database for chemical entities, is still manually created by the Chemical Abstracts Services (CAS) as part of the American Chemical Society. The CAS Registry, as addition to the CAS database, was already introduced in 1965 to overcome problems with identifying chemical entities based on their names. And indeed, CAS still spends a tremendous amount of funding in the manual abstracting and indexing of journal articles, conferences, patents and many other research publications in the chemical domain. For each chemical entity approximately three Euros have to be spent to fully store relevant information in the CAS registry, when extracted from literature and correctly drawn by a domain expert for a structure database. Currently CAS registry comprises over 50 million of substances; however, access is strictly limited to subscribers.

Considering the spirit of open access journals it seems questionable to rely only on high priced commercial abstracting and indexing databases like Chemical Abstracts. Currently there are 111 chemistry journals listed in the Directory of Open Access Journals (DOAJ[9]). But opening up the knowledge of these sources to prac-

[7] http://old.iupac.org/inchi/

[8] http://daylight.com/smiles/

[9] http://www.doaj.org

titioners in the chemical domain requires domain specific tools for searching and (automatically) indexing information. The idea of building chemical databases poses many challenges, the most important being entity extraction, representation and matching.

The problem of entity extraction from full texts for automatic indexing is currently considered for a variety of domains. In chemistry the only open source chemical entity recognition tool currently available is the OSCAR framework [69], which can identify and extract multiple name variations of chemical entities. In combination with name-to-structure algorithms these entity names can be transformed into chemical structure information [70]. Of course the automated recognition of chemical entities is still dealing with the challenges of ambiguity. But, as we will see later, indexing with automatically extracted phrases can already provide sufficient retrieval quality for most documents.

For the internal digital representation and exchange of structures several text-based formats have been developed. Based on the algorithms developed by Morgan [71] and Gluck [72] it is possible to store two-dimensional atom-bond structural representations of chemical entities in a tabular form, so-called connection tables. Besides, linear notations have found widespread use. The early Wiswesser line notation (WLN) [73], or the later SMILES [74], ROSDAL [75] and SYBYL line notation [76] are representations of chemical structures in the form of a linear string of alphanumeric symbols. The latest development is the InChI Code, an open standard for chemical structure description, by the IUPAC [77].

Beside exact substance matching via text strings today's databases have to store chemical structures in several other ways to enable also substructures, or similarity searches. Besides the entire chemical structure saved as a colored, undirected, cyclic graph, fragmentation codes, fragment, or substructure keys and molecular identifiers are used [78], [79]. Fragments are often stored as fingerprints coded as bit vectors. Both structure and substructure search in databases are based on graph isomorphism algorithms. Algorithms and concepts slightly differ by vendor and are mostly proprietary. Here, the general problem is that for each implemented structure database the fingerprints may severely differ. Thus, it is impossible to simply crawl the information from the Web to build up a comprehensive search index.

In current systems these efforts resulted not only in the storage and display of graphical representations of chemical entities, but also in a graphic-oriented search process. It allows a domain expert to actually draw a compound or key fragment as query input. But such specialized information retrieval interfaces are no longer limited to high priced commercial databases in a client-server environment. Recently chemical information about millions of compounds has been made available on the Web. Databases like PubChem[10], Chemspider[11], ZINC[12], ChemBank[13] or

[10] http://pubchem.ncbi.nlm.nih.gov/

ChemDB[14] provide detailed information about some chemical structures, names and properties, also embedding graphic-oriented query interfaces for searching for chemical entities into browsers. But these platforms still require a domain specific indexing and storage of the chemical information in a structure database. A straightforward keyword-based access like provided by common search engines such as Google or Yahoo!, is still insufficiently supported for Web pages dealing with chemical information.

A promising approach for a chemical search engine is Harvard's QueryChem Portal[15]. It allows searching the Web based on an expanded query automatically generated from any chemical structure drawn in a graphical user interface [80]. Similar to our approach, first the chemical structure is converted into a SMILES code which in turn is used for a reference lookup in chemical Web databases like PubChem, ChemBank or Zinc. The lookup provides corresponding synonyms which are then used for a Web search via the Google API. Although such a query expansion definitely is a first step, this approach can only rely on data already correctly indexed by Google. Since most chemical documents are hidden in chemical digital libraries, they still are not retrieved, even by an expanded query. Hence the key to solve this problem lies in proper indexing. Nevertheless, the problem of query overspecialization is still not solved.

In the course of this thesis, we developed the Virtual Library of Chemistry [81] embedded into the chemical search portal chem.de[16]. The chemistry portal acts as a central point of access to various resources from the three major institutions responsible for information provision for chemical research in Germany – the German National Library of Science and Technology (TIB), the Chemistry Information Centre (FIZ CHEMIE) and the German Chemical Society (GdCh). Amongst others these joint services include searching in bibliographic databases, chemistry databases containing comprehensive factual data about molecules and reactions, and full texts of research reports. This ViFaChem portal will be used as use case scenario throughout this thesis.

[11] http://www.chemspider.com/

[12] http://zinc.docking.org/

[13] http://chembank.broadinstitute.org/

[14] http://www.chemdb.com/

[15] http://www.querychem.com

[16] http://www.chem.de

Chapter 4

Towards Quality Assessment

In Chapter 2, we gained a good overview about the state of the art in quality assessment of digital libraries and the problem of domain dependent quality metrics. In this chapter, we will illustrate, why the state of the art, has to be extended when working with semantic technologies. Therefore, we will have a closer look into each step of the development of a digital library. The development is a multi-stage process based on the information life cycle (see 4.1) containing the following steps: Preprocessing of underlying documents (see 4.2), semantic metadata enrichment (see 4.3), indexing of collections and metadata (see 4.4), and personalized document retrieval (see 4.5). During each step quality issues can rise and the reader will be sensitized for them.

4.1. Where and When – Identifying Quality Indicators

For the development of a quality model, it is important to identify when and where quality indicators can be measured, assessed and improved. As we have already shown, information in digital libraries is mainly carried by the metadata of digital objects and the objects themselves. Their life cycle can range from years to centuries in length. Thus, we can reuse the life cycle of information (see Fig. 5), resulted from the research workshop Social Aspects of Digital Libraries [82], to illustrate context. The authors identified three major stages of activity (*creation*, *searching* and *utilization*), each with multiple steps of information handling and processing.

1. Within the *creation stage* authors produce new information (e.g. generating new (meta-)data) or reuse other information to create something new (e.g. combining several data sources). Also the organization and indexing of information is included in this stage.
2. The *searching stage* was identified as semi active because information may only be used periodically. Storing, retrieving, distributing, and networking are included here.
3. The *utilization stage* contains accessing and filtering of information. It was identified as inactive, as information may lie passive over a long time. In this stage, decisions are made whether to retain the information or to discard is as no longer useful. If kept, the information would remain accessible for mining and filtering.

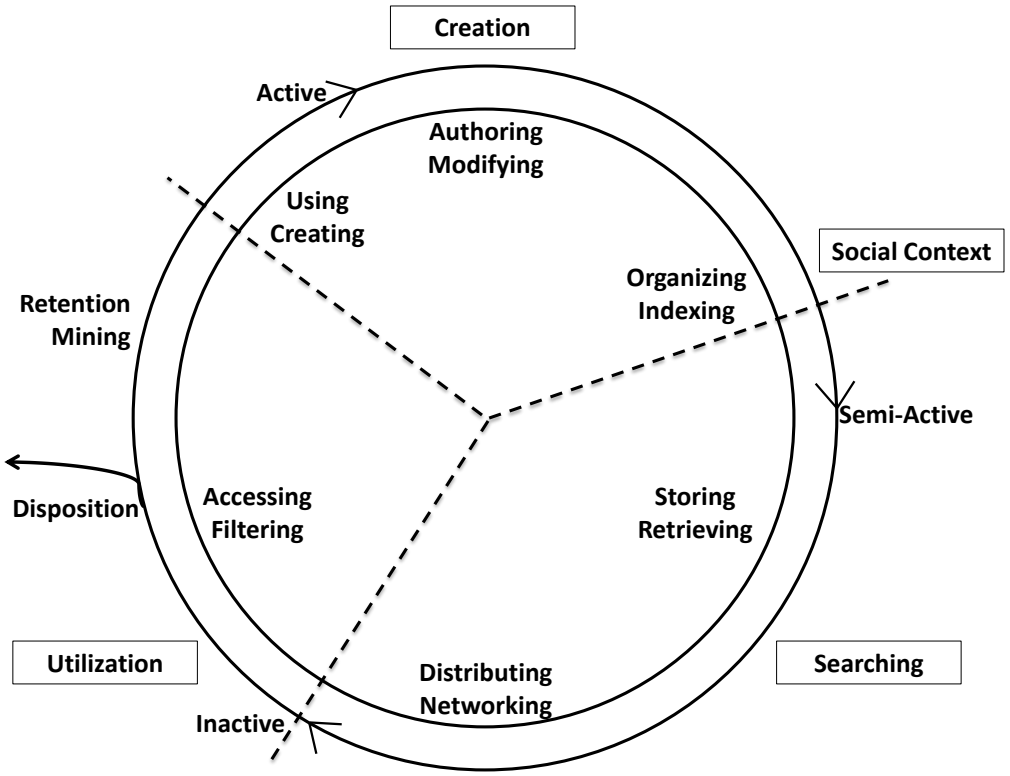


Fig. 5. Information life cycle, reprinted from [83]

Each step in the cycle is influenced by its local context. Thus, we have adapted the general model to our use case scenario, i.e. digital libraries. Based on the information life cycle, our workflow (Fig. 6) comprises the following steps which will be further explained in the following subsections:

I. Convert various file formats and layouts into a single interface representation of the document.
II. Identify all domain entities by entity recognition technologies.
III. Extensive metadata enrichment for all retrieved entities.
IV. Generation of document and metadata indexes.
V. Storing all data in the respective storage engine.
VI. Make metadata indexes public.
VII. Linking index pages to the original sources.

In the following subsection, we will discuss several steps in more detail to sensitize the reader for quality problems and to find solutions for particular problems.

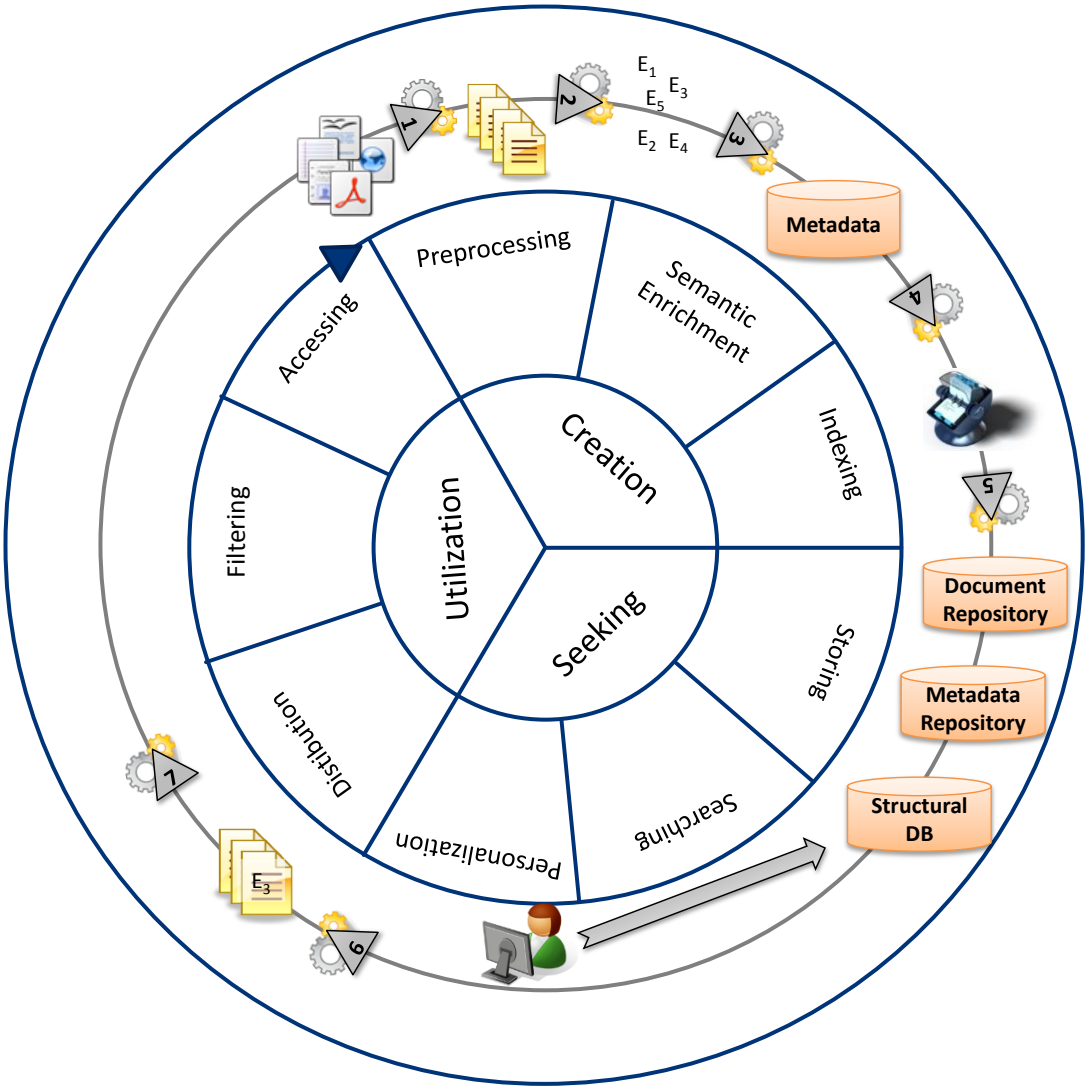


Fig. 6. Workflow Overview

4.2. Preprocessing

Nowadays, entity centric search is an essential requirement for various domains: architects search for specific models, mathematics search for terms and definitions and chemist for chemical entities. Therefore, an important part of document preprocessing is the entity recognition and the preceding document conversion. This conversion is not standardized and has to be adapted for each digital library using its own metadata format, e.g. the TIB uses the so called FTX format. Whereas it is rather trivial to convert structured document formats, e.g., XML or HTML into a well-structured interface representation, the reality is different: most open access journals have only a PDF document collection.

The PDF format is a vector-based page description language that allows for free scalability. The internal representation of a PDF document consists of an absolute position of each letter and graphic element. Therefore, they do not lend themselves to content extraction. Paragraphs and other text elements are not added within a text logically, but whenever a change in position or a line break occurs. In general the probability that names are split into several parts by the OCR process is rather high. Thus, entity extractors have a hard time figuring out whether different parts belong to the same entity or are entities in their own right. Imagine the chemical name *4-(aminomethyl)cyclohexamine* separated into *4-aminomethyl* and *cyclohexamine*.

But especially scientific publications have a high complexity. That way, blocks of text are frequently disrupted by figures, photographs or tables and superscript and subscript numbers and letters (e.g., molecular formula, quotes) are included in the text. Identifying these elements with an acceptable error rate is still a largely unsolved problem. The frequent occurrence of a two-column layout, interrupted by graphical representations, in scientific journals also complicates the automated processing. In Fig. 7 (left) a typical page is presented. On the right is the color coded model of how the actual logical sequence of text blocks is messed up by mechanical processing. The order of the extracted text blocks is understandable for people, but difficult to determine through the OCR.

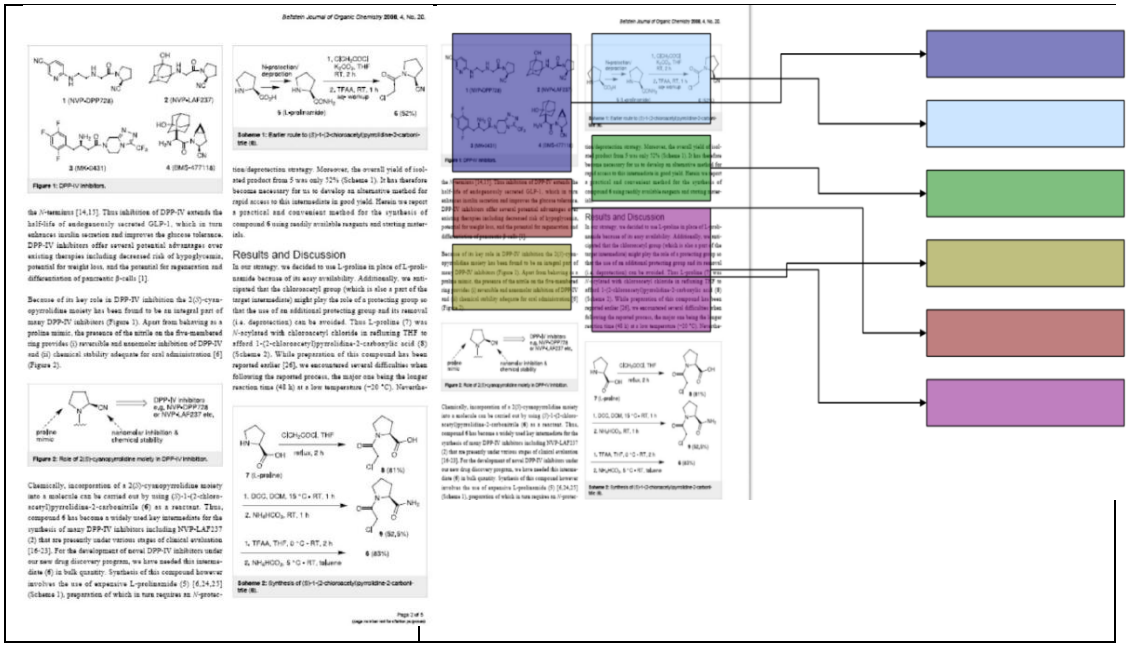


Fig. 7. Processing of a PDF-document

The resulting quality is even worse if the document was digitized using optical character recognition (OCR) software, because the digitized result will always contain additional OCR errors. However, these errors are insignificant when building up a full-text index since standard information retrieval (IR) techniques are not

really affected by unsystematic (OCR) errors. But considering entity-centered domains it might already be an interesting factor in the process of tokenization and entity recognition, also affecting the overall retrieval quality [84].

4.2.1. Evaluation: Influence of Data Formats

We analyzed how data formats influence the results of an automatically entity extraction process. Here, we were not interested in the quality of the extraction process, e.g. if all chemical entities were extracted correctly, but only in the fact if there are differences in the output of the used semantic technique with respect to the used data format. The formats we used are PDF, XML and scanned PDF. The idea is that for each document the same chemical entities have to be found independent of the document format.

For this experiment we took a random sample of around 100 documents from the Beilstein Journal of Organic Chemistry[17]. The documents are available for download in PDF and XML format. As a third format we used a scanned representation of the printed PDF documents. To *automatically* extract the chemical entities we used the OSCAR [69] framework. Therefore, we had to convert the PDF and scanned PDF files into a native text format. For the original PDF documents we used Abode Acrobat to save the documents as text files. The scanned PDFs were processed by an OCR software tool (OmniPage Professional Version 17) to create a textual representation. Finally, we had three different representations for each document which were processed by OSCAR to extract all chemical entities. To determine the ground truth we also manually annotated all documents.

We use a very simple metric to evaluate the quality. We take the number of extracted entities and compare it to the number of entities that can be found in the PubChem database. We assume that entities found in PubChem (entities with structure) have been correctly identified. Fig. 8 shows the total number of distinct entities found in each representation. In the PDF representation 31920 and in the scanned version 36663 chemical entities were extracted. Using the same documents in XML representation only 18262 entities were found. Our ground truth contains 14557 entities and of course all of them have structural information.

[17] http://www.beilstein-journals.org/bjoc

Fig. 8. Number of entities extracted from same documents with same technique but different file formats

Furthermore, we visualize the ratio of entities with and without structural data (Fig. 9). The figure shows that the recognition rates drastically vary with respect to the data format used for extraction. The highest number of entities contained in PubChem is found for the scanned PDF files. Actually this result does not mean that scanned PDFs deliver the best quality. Many of the found entities can be lead back to recognition errors, like e.g. the chemical entity *C* which can be a fragment from a table or a figure. But the experiment shows that even for a very simple metric the output strongly varies with respect to the input format of the actual document.

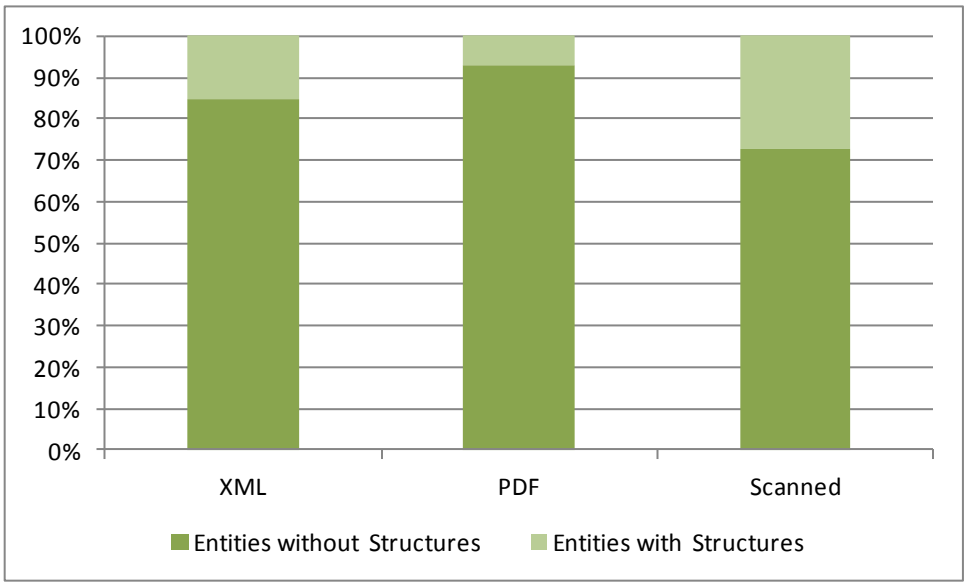


Fig. 9. Ratio of entities with and without structural information

4.2.2. Conclusion

Illustrated for, but not limited to, the domain of chemistry, we could show that even in the conversion of documents, a lot of errors can occur. In particular these errors are worse, because they will be propagated to the following steps of the workflow, e.g. entity recognition techniques. These results lead to the assumption that it will not be possible to consider the quality of the document conversion independently of following steps, e.g. entity extraction. It is obvious that the overall quality will be a cumulative measure that is calculated over the entire workflow. Therefore, it is very important to know which kind of document format the original source has and to have metrics to judge the correctness of such a conversion step. Again, we can also see that this metric is probably very domain and task specific. Thus, in a final quality model it is necessary to store all information regarding the development process.

4.3. Semantic Metadata Enrichment

After the document preprocessing, i.e. the document conversion into a system specific interface format, it is now possible to extract metadata from the document corpora. As we have shown, the quality of the extracted metadata will be dependent on the quality of the preprocessing, e.g. it is not possible to fix OCR errors and thus metadata extractors will have lower quality. But still, the metadata enrichment is very important for digital libraries. Since metadata can express concepts not explicitly occurring in the document, (or leave out concepts explicitly mentioned, but not relevant for the document) the use of a metadata index generally leads to better precision and recall in information services. Hence, digital libraries provide an added value over unstructured document collections by offering meaningful access paths.

Today, the description of a document is based on four types of metadata. It must be distinguished between the conventional pillar, i.e. bibliographic metadata like author, title, and date, and semantic metadata, which describe the content of a document and allow for a classification in the context. This metadata can be cross-disciplinary, for example by means of a universal classification like Dewey Decimal Classification (DDC[18]), or discipline dependent. In the context of the TIB this can be the Mathematics Subject Classification (MSC[19]), Chemical Entities of Biological Interest (ChEBI[20]), or ACM Computing Classification System[21]. In order to achieve a much higher quality in the automatic indexing and retrieval of scientific literature, other criteria come into play in the form of secondary decision support. In particu-

[18] http://www.oclc.org/dewey/

[19] http://msc2010.org/Default.html

[20] http://www.ebi.ac.uk/chebi/

[21] http://www.acm.org/about/class/1998

lar referential metadata, citation analysis or typical Web 2.0 techniques are included into the analysis. Fig. 10 illustrates the 4 pillar model of document metadata. For each pillar, several techniques are available and we showed in Chapter 2 that for some of them quality measures are already available.

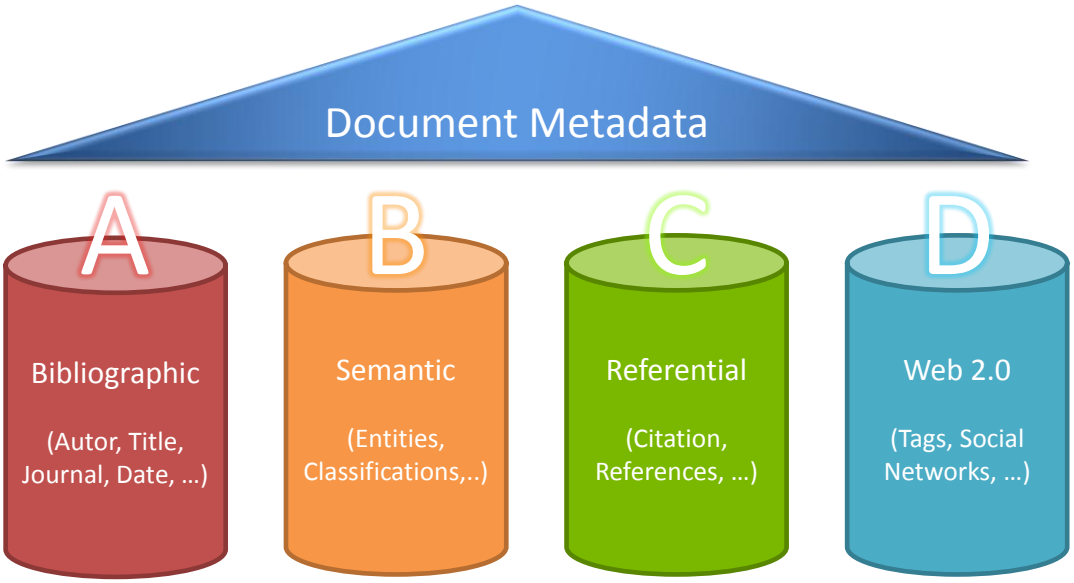


Fig. 10. 4 Pillar Model of Document Metadata

The potential of the usage of semantic information has been demonstrated especially using manually maintained ontologies in various fields. For example, in the field of medicine, biochemistry or bioinformatics the use of the Medical Subject Headings (MeSH[22]) facilitates the development of advanced portals such as GoPubMed[23] or Q-Sensei[24]. These portals offer new ways to classify, rank and cluster documents: metadata is associated with a thesaurus, a taxonomy or ontology.

However, given the exponential increase in newly published items even for focused collections, librarians face two serious problems. First it is increasingly costly and time consuming to properly index new items (leading to a delay in actually offering the item to customers); second in an ideal collection, the indexing has to foresee all possible (future) uses for a specific item. Moreover, the information overload for the individual customer and the increasing specialization of (research) interests force indexes to be more and more specific in the choice of appropriate indexing terms. In fact, the vision of today's digital libraries is to provide *personalized information spaces* for each individual customer.

[22] http://www.nlm.nih.gov/mesh

[23] http://www.gopubmed.org/web/gopubmed

[24] http://www.qsensei.com

To this end, semantic technologies have been recently proposed to bring a higher rate of automation into the indexing process. In essence semantic technologies rely on statistical methods to assess textual documents and to some degree are therefore capable of mining 'hidden' information from collections. The advantage is twofold, first document processing becomes less expensive and a higher degree of personalization is possible. Though, due to the nature of statistical methods, using these semantic techniques may not result in the same retrieval quality as manual crafted metadata. Second, for libraries, this potential decrease in quality is a serious concern; if users cannot trust in the results, the added value over simple Web searches becomes questionable. Hence, before a specific semantic technique can be adopted for use, libraries need a way to gauge the impact of the technology's use in the retrieval process. Even better would be the possibility to gauge the impact of a class of semantic techniques, e.g. classification techniques. But especially for semantic metadata the quality of the filled-in values may differ a lot, depending on the interpretation of the semantic meaning of a metadata field, for different consumers which is illustrated in the following experiment.

4.3.1. Semantic Meaning of Metadata Fields

In this experiment we show that it is important to know the semantic meaning of a metadata field. Therefore, we analyzed chemical reactions occurring in our repository. We crawled 2300 documents from the Archive of Organic Chemistry (ARKIVOC)[25]. We defined a metadata field that should include *chemical reactions* found in the respective document.

We used three different approaches to extract the required metadata (the results are shown in Fig. 11). Firstly, we used OSCAR to automatically identify all reactions found in the document resulting in around 16500 reactions. Secondly, we utilized the Name Reaction Ontology (RXNO)[26] as a dictionary for named reactions, like e.g. *Heck* or *Suzuki* reaction. We used this dictionary for extracting all named reactions out of the documents resulting in 337 named reactions. Finally, we used a very naïve regular expression based approach searching for the '*-reaction*' pattern resulting in 54 reactions.

[25] http://www.arkat-usa.org/arkivoc-journal

[26] http://www.rsc.org/ontologies/RXNO

Fig. 11. Chemical reactions versus named reactions and reaction pattern

This quantitative difference becomes even more obvious, if we have a further look to the upper bound of possible reactions. In our test corpus we identified around 2200 distinct reactions. We assume this amount as optimistic global upper bound for available chemical reactions and compare it with the 300 reactions defined in the RXNO. It is remarkable that the named reactions are only 12% of all possible reactions (see Fig. 12).

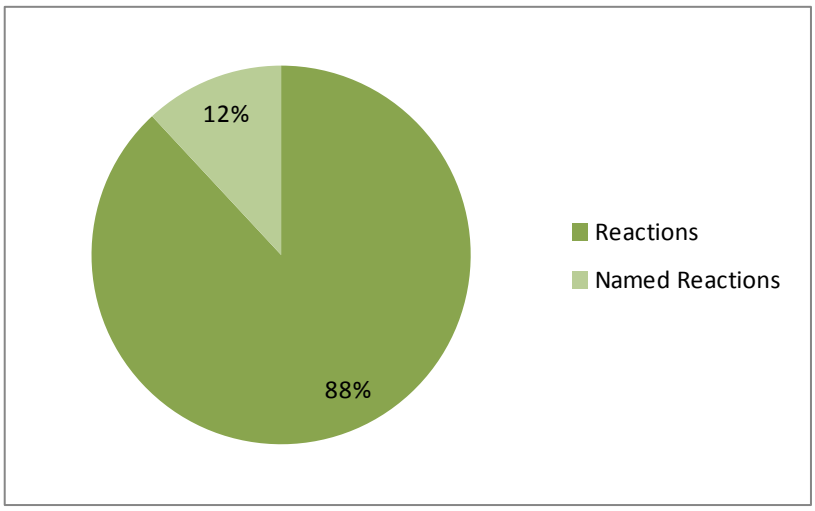


Fig. 12. Distinct reactions found by OSCAR compared to RXNO

Please note both approaches provide suitable inputs for the defined metadata field. However, the results show that the outcome is totally different. Only two percent of all found reactions are named reactions. Although both approaches deliver correct results for the defined metadata field the results highly differ. Considering a library who is interested in showing all related reactions a content provider only offering named reactions is not sufficient. Therefore, it is mandatory to

provide additional information of the semantic meaning of a specific metadata field, respectively of the included information stored in that field. Taking this into account, we will provide a roadmap for developing quality assessment measures for semantic techniques.

4.3.2. Experiments over a Digital Collection of Chemical Documents

We conducted a user study by observing experts, in our case practitioners in the field of chemistry. During the experiment, these practitioners have been working with a topic restricted document collection whose metadata was automatically created by semantic technologies. The aim of the study was first to get a deeper understanding of the process of evaluating metadata and assessing the individual expectations and second the actual helpfulness of the metadata provided.

For the experiments we used a corpus of 1000 documents randomly extracted from the Journal of Synthetic Organic Chemistry published by Thieme Publishers, Stuttgart Germany. For the metadata extraction we focused on the author keywords which were subsequently used for automatically creating folksonomies. The actual graphs were calculated by the Semantic GrowBag technique [44] investigating higher order co-occurrences (known from computational linguistics [85]) of the keywords in relation to the respective documents. A term *A* is considered to be '*more general*' than some term *B*, if *B* usually occurs together with *A*, whereas *A* also occurs in other contexts. In that case a directed edge is added from *A* to *B*. Together with the graph structure the Semantic GrowBag technique also allows to assess the confidence for each relationship visualized by bold (strong) or dashed (weak) arrows. The Semantic GrowBag uses a biased page rank algorithm [86] to determine this confidence.

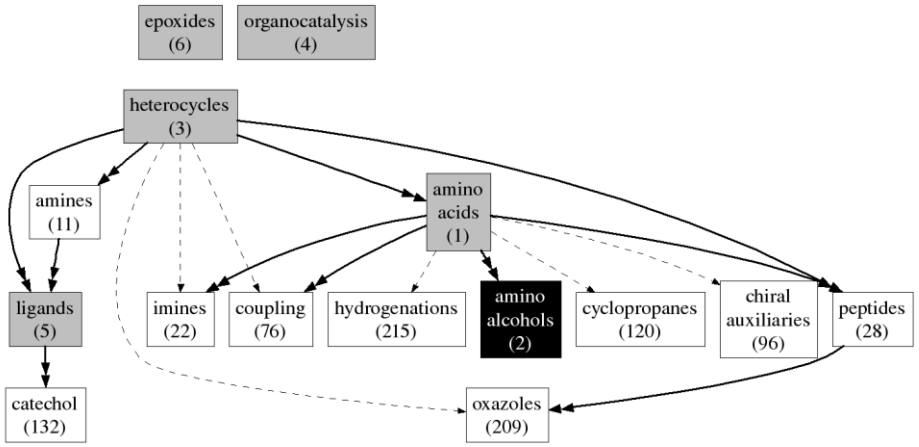


Fig. 13. The generated GrowBag graph for the keyword '*amino alcohols*'.

In Fig. 13 '*amino acids*' is considered more general than '*amino alcohols*' which is indeed justified by *amino alcohols* being a subclass of *amino acids*. Though, please note that a relationship as given by the GrowBag graphs does not always express a

subclass (or is-a) relationship, but just points out that in terms of usage as reflected by the document collection the parent term is more general than the child term.

We extracted a total of 680 graphs, each representing the semantic environment for all sufficiently discriminative keywords. The page rank of each term (the number in brackets) in the graphs was also used to create the related tag clouds for the keywords (e.g. Fig. 14). The respective size of each term in the tag cloud is proportional to the page rank value of the term in the GrowBag graph. Please note that in principle the tag cloud contains all information which is available in the graph (terms and their respective page rank) just the hierarchical structure (edges) is missing.

Fig. 14. The generated Tag Cloud for the keyword *amino alcohols*.

For the actual experiments we randomly chose three query terms for each expert to evaluate the quality of the given graphs and the respective tag clouds. All experts were asked to think aloud after being exposed to the individual graph or tag cloud and provide feedback on how they assessed quality and which metadata items were considered to be sensible for the average user of the respective collection. Moreover, after reviewing the metadata for each query term, the experts were asked about their expectations in terms of organization of the metadata and the respective correctness and completeness of the automatically created metadata vocabulary.

A Case Study

In this case study we describe a typical expert's interaction with a generated graph (Fig. 13) / cloud (Fig. 14) for the query term '*amino alcohols*' to illustrate the conduction of our user study. A first expert was asked about the graph representation and a second about the cloud representation. Please note that graph and cloud contain the same terms and just differ in the visualization and connections between terms.

Given the graph as shown in Fig. 13, the expert immediately pointed out that the query term represents a class of *chemical entities*; therefore, he expected to see several *attributes of this class*, typical *reaction names* where amino alcohols are used, *technical uses* and some specific terms from an *analytic* point of view. Following these expectations he clustered the elements into the following groups:

— reactions: 'coupling' and 'hydrogenations'
— classes: 'cyclopropanes', 'oxazoles', 'heterocycles', 'peptides', 'imines', 'amines', 'amino alcohols', 'amino acids', and 'epoxides'
— general concepts: 'chiral auxiliaries', 'organocatalysis', and 'ligands'
— instances: 'catechol'

In a next step the expert noticed that there are big differences in the generality of the terms, e.g. '*heterocycles*' has been seen as a very general term whereas '*cyclopropanes*' is a more specific term. For the last step of interaction the relationships were analyzed: the expert considers some useful, e.g. '*peptides*' are connected via their building blocks '*amino acids*' with '*amino alcohols*' which fits better than a direct connection to 'amino alcohols' and others not useful, e.g. '*catechol*' which represents a '*hydroxyl benzene*' with no obvious connection to '*amino alcohols*'.

After giving the equivalent tag cloud (Fig. 14) to the expert, it was interesting to note that the way of interaction was to a large degree identical with the graph-based representation: again the expert started with predefined categories and tried to assign the terms, second the generality of the terms was judged and third the terms were linked to the query term. It has to be pointed out that the expert working on the tag cloud had much more problems during the last step, due to the way of visualization. For instance, he was surprised about the font size of '*cyclopropanes*' and '*oxazoles*'. Due to the fact that '*cyclopropanes*' is not related to the query term, he expected the font size to be much smaller than, e.g., the size of the heavily related term '*oxazoles*'.

Experimental Results

The evaluation of our observations showed that all practitioners made three major steps during the interaction with the offered metadata. All experts started with some initial expectation for the categorization of metadata terms. First, they categorized the query term, e.g. as a substance class and then settled on semantically related subcategories based on the main category. It was interesting to see, that these subcategories slightly varied based on the background of the expert. For instance, an expert in the domain of medical chemistry also mentioned the pharmacological impact, whereas a process engineer mentioned environmental perils and toxicity. This observation leads to the conclusion that a categorization of the terms, as it is done, e.g. for the faceted browsing, is indeed useful for the customers and that the structure of a tag cloud may not always be sufficient for visualizing this kind of semantic metadata. It seems that the distribution of terms over relevant categories is one useful metric for measuring the quality of the generated metadata. In our experiments over 90% of the expected categories were indeed filled by matching keywords.

In the second step the experts tried to understand the content of the graph / cloud. For this purpose they evaluated the terms regarding their respective generality / specificity. This was done without considering the query term. This step has been used by the experts to eliminate outliers in terms of very general or very

specific keywords. In particular during our experiments the experts considered 32% of the provided keywords as being too general / specific for the respective graph / cloud.

The last step was the evaluation of the semantic closeness regarding the query term. During this evaluation step the visualization of the metadata affected the experts. Working on the graph every term was judged individually and depicted relationships were readily taken as explanations. The experts which worked on the cloud did not have these relationships and, therefore, were confused about some terms. Even worse, the font size of the term influenced the experts far more than the confidence in the GrowBag graph. These observations imply the usage of different visualizations: using a cloud for well-connected terms and using a graph for the others. In summary, the experts used their individual knowledge to understand the occurrence of the terms and if they could not make a direct connection between a keyword and the query term, they tried to connect the term via some other occurring terms in the graph. If this also failed, they considered the term as wrong or irrelevant for the query. In our experiments this happened with 12% of the occurring terms: this means that 88% have been classified correctly.

4.3.3. Towards Measuring Semantic Information Quality

The experiments in the previous subsection provide some general ideas regarding the measurement of the quality of a semantic technology. Generally speaking quality can be defined as *correctness* of information. For the field of chemistry this is especially true for data maintained in typical databases like molecular weights or boiling point of substances. However, with respective to semantic, e.g. given by author keywords, the actual correctness is somehow difficult to assess. Observing the expert we found that experts gorge the correctness rather in terms of helpfulness of a keyword and the understandability of the keywords' relationships to a query term. According to the three steps observed during the experiment we found some commonalities between experts. Based on this we will now discuss three preliminary quality metrics that will be further evaluated in this section.

The evaluation of the experiments showed that all experts from the start have an implicit course topic map together with possible classifications for entities in mind. Although the topic map differed with the individual interests of the expert, it is interesting to note that the basic entity classification was very similar (in a way reflecting the typical cognitive instruments of a chemist). According to this implicit classification each expert tried to categorize the metadata terms automatically created by the semantic technology. The choice of categories under consideration slightly differed according to the query term and the experts expected at least the closest categories to be filled with keywords found in the graphs, respectively clouds.

This leads to the *degree of category coverage (DCC)* metric which has to measure how many of the expected categories are actually filled with terms. The more categories are filled the better the result quality is.

With *C* := {*c* | *c* relevant category in the topical classification} we define:

$$f(c) = \begin{cases} 1 & \text{if there is at least a single term } t \text{ in category } c \\ 0 & \text{else} \end{cases} \quad (1)$$

In addition the metric also has to measure how many of the given terms do not fit to at least one of the expected categories. The more terms can be allocated the better the result quality is.

With *T* := {*t* | *t* term from a given metadata subset} we define:

$$g(t) = \begin{cases} 1 & \text{if term } t \text{ belongs to some category from } C \\ 0 & \text{else} \end{cases} \quad (2)$$

This results in:

$$DCC = \frac{\sum_{i=1}^{|C|} f(c_i)}{|C|} + \frac{\sum_{j=1}^{|T|} g(t_j)}{|T|} \quad (3)$$

The *Semantic Word Bandwidth (SWD)* should reflect the results of the second interaction step: the experts estimated the overall generality / specificity of the given terms. Of course this bandwidth can only be evaluated with respect to the highest possible bandwidth. The smaller the bandwidth the more focused is the set of related keywords.

Considering categorizations where we can rely on some is-a hierarchy, e.g. taxonomies of chemical substances, it is quite simple to determine the bandwidth. In this case we have to identify the depth within the hierarchy for each term. Using the maximum and the minimum depth of terms normalized by the total depth of the hierarchy (*maxdepth*) the semantic word bandwidth can be defined as follows:

$$SWB = \frac{\max_{t \in T}(depth(t)) - \min_{t \in T}(depth(t))}{maxdepth} \quad (4)$$

In cases where no is-a hierarchy is given it is much more complex to estimate the semantic word bandwidth. For instance, considering substances (e.g. reactants or catalysts) involved in chemical reactions could be considered more specific but in any case this would need a complex ontology describing the relationships for reactions which can currently not be found in the market place.

The last measure used by the experts tried to determine the usefulness of a term in relation to the query term, named *relevance of covered term (RCC)*. If we

consider again some is-a hierarchy or an ontology we may express the usefulness of a term in relation to a query term as the semantic similarity between those terms. The total relevance can then be established as the average similarity of keywords to the query term.

Practically this can be done by analyzing the underlying ontology. All keywords are associated with concepts in the hierarchy. A direct method for measuring the respective similarity is finding the minimum length of any path connecting the two concepts [87]. However, according to [88] this may not be sufficient for more general and larger ontologies and thus the similarity should be a function of the attributes path length, depth and local density.

Another possibility to measure the relevance of the covered terms may be reflected by using independent semantic techniques. In our example the Semantic GrowBag uses statistical information to compute higher order co-occurrences of keywords. Thus, the relations shown in the graphs reflect some characteristics of the underlying document collection. The naive way of interpreting the results is that all terms covered by one graph are somehow used together with the query term. If we assume that terms which are more related to the query term are also generally used more oven in relation with some document, this should also be reflected by a simple Web search query. Thus, a two term query for a query term *qt* and a word w_1 which are closely related should result in more hits than a query for *qt* and some word w_2 that are not as closely related. Preliminary experiments based on our used graphs seem to support this assumption, e.g. a Google search for the query '*amino acids* AND *amino alcohols*' yields 39,800 hits and the query '*amino alcohols* AND *cyclopropanes*' only yields 2,540 hits.

4.3.4. Evaluation of Quality Measures for Semantic Techniques

We conducted a user study with domain experts, in our case practitioners in the field of organic chemistry, for evaluating the three metrics DCC, SWD and RCT. The aim of the study was first to get a feeling whether the defined metrics are useful and second how important the information provided is compared to classical forms of visualization.

For the experiments we used a corpus of 4554 documents extracted from the Journal of Synthetic Organic Chemistry (SYNTHESIS)[27] published by Thieme Publishers, Stuttgart, Germany. For all papers we extracted the author keywords (9554) and eliminated all those with little discriminating power (terms like '*experiment*' or '*synthesis*') occurring in many papers. For the remaining set of about 1600 terms folksonomies were generated using the Semantic GrowBag technique [44], which relies on higher order co-occurrences of keywords in relation to the respective documents. For the actual experiments we randomly chose ten query

[27] http://www.thieme-chemistry.com/en/products/journals/synthesis.html

terms for each expert to evaluate the quality of the respective folksonomy. For each query term we generated three different kinds of visualization: the original GrowBag graph (Fig. 15, query term in black box), the respective Tag Cloud (Fig. 16) and a concentric circle diagram (CCD) (Fig. 17, query term in the center).

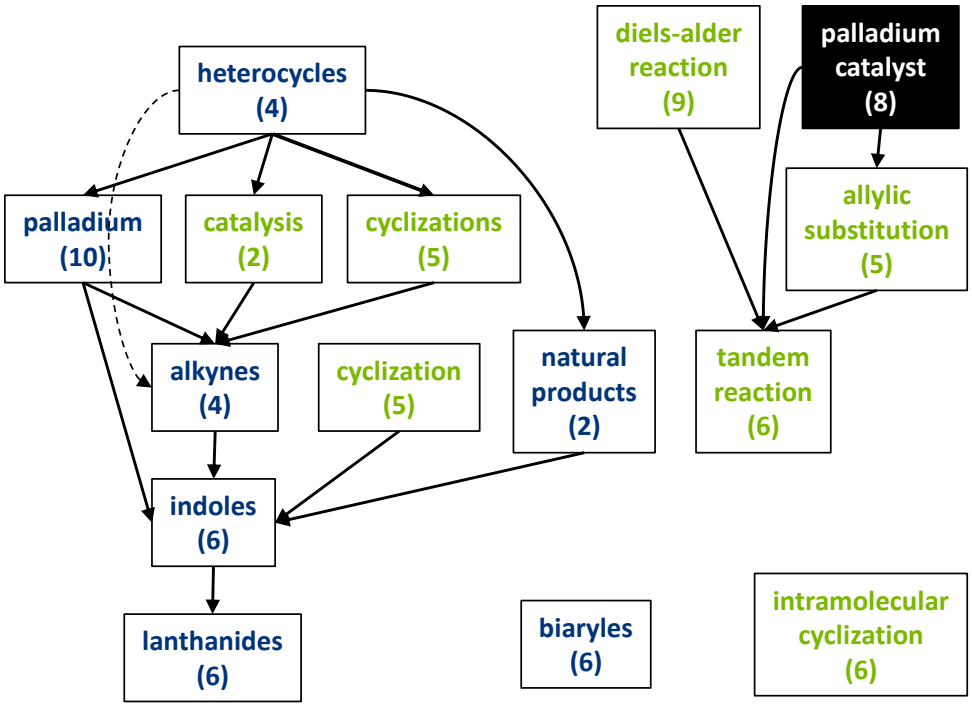


Fig. 15. The generated GrowBag graph for the keyword *palladium catalyst*.

Fig. 16. The generated Tag Cloud for the keyword *palladium catalyst*.

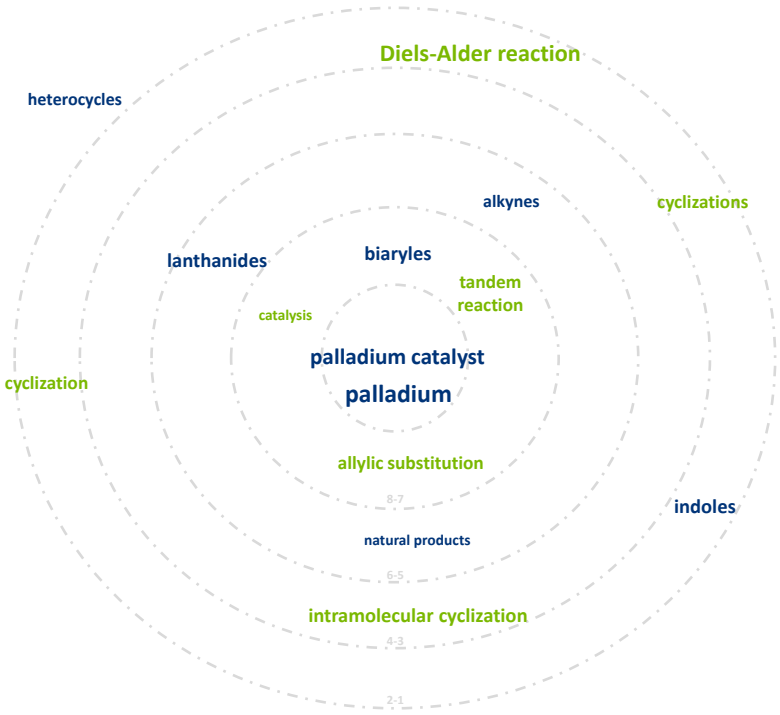


Fig. 17. The generated concentric circle diagram for the keyword *palladium catalyst*.

Basically the information provided by our three quality metrics, i.e. related category, overall specificity topical and distance to query term, can be represented by the visual features text color, text size and spatial layout. Please note this information can be easily visualized in the CCD, whereas the other two visualizations lack the possibility to visualize the distance in an intuitive way. For the tag cloud we tried several clustering algorithms and visualized the terms in clusters, thus sacrificing the compactness of display for the possibility to show spatial relationships. However, since intra-cluster similarity usually clashed with the individual terms' relationship to the query users tended to be confused by that notation. Hence, in our experiments we tested the advantages of compact visualization versus the benefits of the information provided by the distances to the query term.

Designing the User Study using the Semantic GrowBag Technique

To control the environment we ensured that all participating domain experts were recruited from the field of organic chemistry and in particular familiar with the focus area of the SYNTHESIS journal. Hence, we could expect only slight variations with respect to the individual knowledge spaces. Although the experts did not know about the specific Semantic GrowBag technique used for deriving the graphs, all participants had prior experience with the use of ontological information retrieval and were proficient in using computing devices.

As stated above for all users we randomly selected ten query terms and confronted them with the three different visualizations in individual questionnaires. Filling in the questionnaire took only about half an hour of time. To illustrate the design of our user study let us focus on an example evaluation workflow for the query term '*palladium catalyst*'. The respective visual representations are shown in figures 15 – 17. Each questionnaire was divided into three major blocks:

- The first block of questions in the questionnaire focused on the first impression with respect to the diagrams. Users were asked to rank the different diagram forms for each query term regarding the intuitive understandability, i.e. the degree of ease to grasp the concepts contained.
- After evaluating the first impression the second block should prove if the users intuitively interpreted the diagrams in the correct way. Therefore, each metric and the correlation between the metrics' outcome and the diagrams were explained. With this knowledge the users were again asked to rank the different diagram forms.
- The third block actually measured the correctness of the three quality metrics. Therefore the domain experts were asked to rank the visualized metrics' outcome for each query term.

In more detail, the metrics explained in the second part of the evaluation concluded in the following visualization. Focusing again on the example query term '*palladium catalyst*', all terms can be categorized into the two categories, i.e. reactions (light) and chemical substances (dark). The size of each keyword represents the overall specificity, e.g. '*Diels-Alder reaction*' is a quite specific term as it represents a concept of specific reaction with given reactants, products, solvents and reaction conditions. This way a domain expert reading the term 'Diels-Alder reaction' - maybe in connection with a substance - has a good impression of a reaction scenario and possible products. In contrast, the '*tandem reaction*' is an unspecific term describing a broader concept of a reaction type with much more space for interpretation. The term '*tandem reaction*' just defines a cascade of reactions from an educt to a product without the isolation of any intermediate product: the actual reactions of the cascade are not defined in detail by this concept. As already stated above, the closeness of a term in relation to the query term can only be visualized within the CCD, i.e. the distance of each keyword to the circles' center. Thus, the query term '*palladium catalyst*' is located in the circle center. Closely related terms like e.g. '*palladium*' and '*tandem reaction*' are located nearby, whereas only loosely related terms are located far away e.g. '*heterocycles*'. The closeness of '*palladium catalyst*' and '*palladium*' is obvious as a palladium catalyst contains the metal palladium, which in turn defines the functionality of the catalyst. Tandem reactions are most often catalyzed reactions with a high stereo selectivity induced by various classes of palladium catalysts. An example for a loosely related term is '*heterocycles*', which represents a general concept of a substance class with a rather weak relation to the term '*palladium catalyst*'.

For the last part of the experiment the domain experts were given a scale divided into five degrees (0 - 4) of satisfaction (see Table 4).

Table 4. Evaluation scale for part 3 of the experiment

Value	DCC: percent of occurring concepts	SWD: percent of matching proportional font sizes	RCT: percent of matching distances
4 - completely satisfied	> 90%	> 90%	> 90%
3 – mostly satisfied	~ 75 %	~ 75 %	~ 75 %
2 – satisfied	~ 50%	~ 50%	~ 50%
1 - partially satisfied	~ 25%	~ 25%	~ 25%
0 – unsatisfied	≤ 10%	≤ 10%	≤ 10%

Experimental Results

In the first part of the experiment we evaluated the first impression and the intuitive understandability of the respective visualizations. We expected a high rank of the tag cloud as it is a compact and well known kind of visualization. Surprisingly, as can been seen in Fig. 18 the concentric circle diagrams (CCD) were already ranked considerably higher immediately claiming about 95% of the position one ranks with an average rank of 1.07. In contrast the tag cloud visualization just got an average rank of 2.1 and the remaining 5% of position one ranks. The (somewhat harder to understand) ontology graph was never ranked at position one and only got an average rank of 2.82.

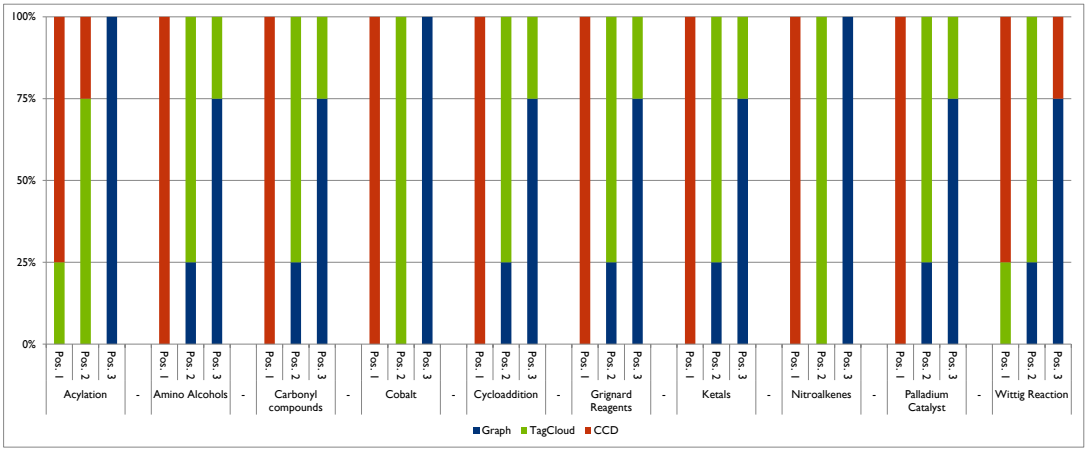


Fig. 18. First impression results

It is interesting to note that in topically focused document collections quality information blended into navigational information or categories is indeed attractive for users. This also shows that even the simple CCD is already an intuitive way of visualization for our quality metrics. Also a later interview with selected domain experts confirmed this: they explained the lower rank of the tag cloud, because the co-occurring terms were sometimes misleading. But the adoption of the distance in the CCD clarified intuitively that some terms may belong to the query term in a rather loosely coupled way.

In the next step of the experiment, the visualizations' semantics in terms of encoded quality measures was explained in detail and the domain experts were asked to re-rank the three kinds of visualization. As expected, the CCD was still most often ranked at position one (see Fig. 19). However, a marginal loss of 2.5% in position one ranks occurred, still resulting in 92.5% of top positions and an average rank of 1.1. At this stage, although gaining 7.5% of the overall position one ranks, tag clouds experienced a slight drop in the average rating (2.15). This can particularly be attributed to their limitations becoming clear during the explanation of the semantics: users better understood their power of compact representation, but also their difficulties in discriminating terms. Again, the graph representation was never ranked first but the re-ranking still resulted in a slightly better average rating of 2.75. Further interviews with the domain experts have shown that they liked the tag cloud more than the CCD in situations, when confronted with very sparse CCD diagrams. This shows that there is a tradeoff between compactness of the visualization and the transported information. One possibility to handle this tradeoff would thus be a digital library interface where first a tag cloud of the digital collection is shown and once a user selects a term for deeper investigations, the respective CCD is shown.

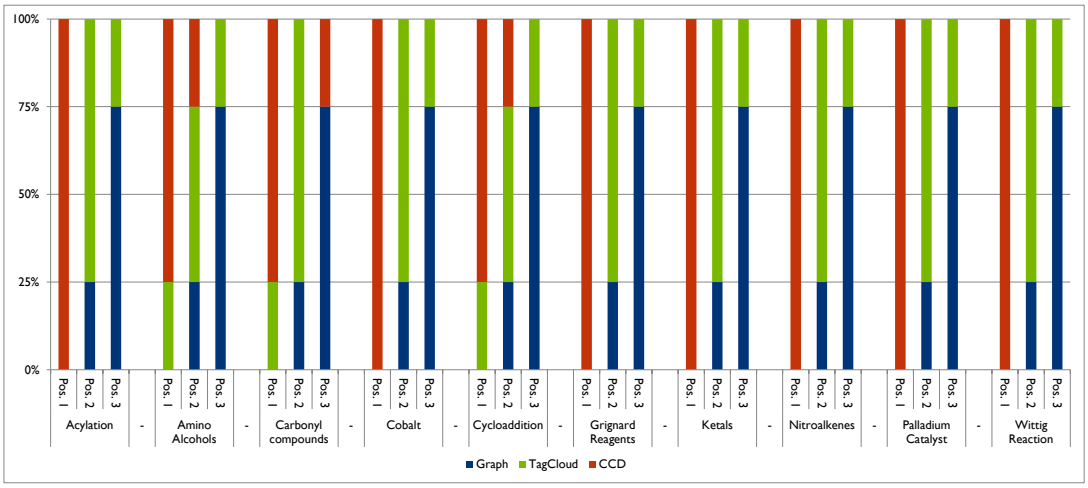


Fig. 19. Second impression results

Due to the fact, that the CCD is the only diagram which represents our metrics' entire outcome and that the rank has only slightly been decreased after explaining the quality metrics contained, our three metrics indeed seemed reasonable for the domain experts. A deeper investigation as last part of the evaluation further substantiates this prediction. We asked all experts to consider the three metrics individually and evaluate the terms provided by the Semantic GrowBag technique for the ten test queries. As can be seen in Fig. 20 on a scale from 0 to 4 none of the metrics' outcome has been ranked less than 2, i.e. the 50% mark of satisfaction. In average the degree of domain coverage (DCC) was ranked with 3.20, the semantic bandwidth (SWD) with 2.82 and the relevance of covered terms (RCT) with 3.18. On average the domain experts were mostly satisfied with the quality of the Semantic GrowBag technique's generated metadata and also the proposed quality metrics' usefulness in quality assessment was confirmed.

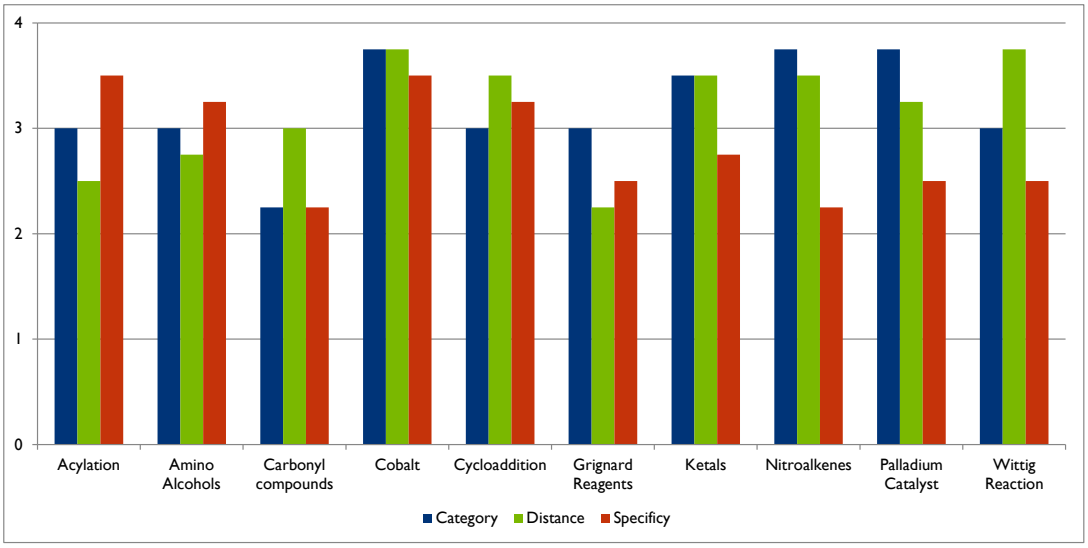


Fig. 20. Rating of the correctness of the quality aspects from unsatisfied (0) to completely satisfied (4)

4.3.5. Conclusion

Semantic techniques are ubiquitous in modern information systems and digital collections. In this section we dealt with the question whether the expected loss of quality due to the use of statistical techniques can be measured. We argued that the development of such measures is especially important for their safe and sustainable application in digital libraries which generally have higher quality constrains in comparison to, e.g. Web search engines. Putting the focus on automatic metadata creation as provided by related keywords we conducted a user study in the field of chemistry observing some experts' interaction with the created metadata. The study resulted in three major observations:

1. Domain experts always started from a (reasonably similar) cognitive classification of possible entities. They expected to find relevant terms with respect to all expected classes.
2. Considering the given metadata all experts expected to find a similar degree of generality / specificity of the keywords. The respective degree was derived relative to the general understanding of the respective domain.
3. Assessing the type of relationship between each keyword and the query term all experts tried to embed the terms in a common context. With increasing broadness of the context, the satisfaction with the keywords decreased.

Based on these observations we proposed three measures namely degree of category coverage (DCC), semantic word bandwidth (SWB) and relevance of covered terms (RCT). Furthermore, we investigated the usefulness of the proposed metrics and their information gain. For this purpose, we conducted a user study with domain experts. We showed that it is indeed useful to measure the quality of a semantic technique: the domain experts were easily able to assess the outcome of the technique and gained insights into what quality to expect during their information gathering. The Semantic GrowBag, a statistic technique relying on term-co-occurrences for deriving metadata, was graded with an average of about 'mostly satisfied', i.e. about 75% complete, related, and specific. Moreover, by also providing the quality values visually for each term within the navigation elements, domain experts were less confused (especially when interacting with a low grade folksonomy).

For the investigation during our survey we used a very simple kind of diagram derived from fisheye view interfaces visualizing quality values of each term by its distance to the query term: the Concentric Circle Diagram. Still, we were able to show that users are more satisfied by the experience of using this kind of diagram than the semantically rather shallow, yet popular tag clouds. This implies that the underlying quality should not only be used within the creation process of digital corpora but should also be used during the information seeking process of a user within the collection.

4.4. Indexing

In the previous section, we discussed quality issues during the metadata enrichment of already preprocessed documents. This metadata is a substantial part of the information seeking process in a digital library. By indexing also metadata and not only full texts, the service quality can be significantly increased because documents can be better indexed and topical sorted. But how can the quality of an index be evaluated?

In general, the quality of an index can be evaluated by means of the accurate representation of a document within the index whereas the effectiveness of an index measures whether an indexed document is correctly retrieved every time it

is relevant to a query [89]. It is also popular to determine the quality of an index by the *interindexer consistency*, i.e. the extent to which agreement exists between indexers on the terms to be used to index the same document [90], [91]. These measures will lead to problems while talking about semantic metadata enrichment especially talking about synonyms because the terms which can be used for document description dependent on the applied semantic technique. Thus, measuring the index quality seems to be a double sided process:

1. Measuring the quality of underlying semantic techniques
2. Measuring the quality of the index itself according to the above measures

Furthermore, digital libraries have to face the *problem of the hidden Web*. That means, most of the indexes build by digital libraries are not detectable by standard Web search engines and thus are hidden to most users because this is the standard way of information seeking in the Web. Hence, only domain experts, knowing the topical relevant collections, are able to find *all* relevant information. Consequently, the quality of an index should be extended to also measure the precision and recall achievable for a specific collection by using a standard Web search engine. However, consumers have different workflows and expectations when searching for relevant literature, strongly depending on the scientific domain, the level of expertise, and the task at hand. Thus, it is also necessary to compare the performance in the sense of precision, recall and speed of standard search functionalities and Web search engines. The basic idea of our approach is to automatically create and link enriched index pages comprising metadata for a document collection. By linking these pages to the original documents or the digital library entry, they serve as a search index over the hidden index. In the following experiment we instantiated this approach for the chemical domain.

4.4.1. Evaluations for the Domain of Chemistry

In the domain of chemistry information seeking is essentially centered on chemical entities. Moreover, practitioners, as well as academic researchers, are usually interested in finding *all* related documents to individual chemical entities. For both the search is basically recall-oriented because especially for synthesis procedures or production processes missing information about for instance existing patents or expected yields may lead to considerable financial losses.

The usual representation of chemical entities is based on chemical structures which are embedded (as images) into the documents. Whereas domain experts can easily identify the shown structures and classify them in the context of the document, it is currently impossible to extract this information automatically. First commercial tools like CLiDE Pro[28] or chemoCR[29] show the basic desirability.

[28] www.keymodule.co.uk/CLiDE.html

[29] www.scai.fraunhofer.de/chemocr.html

However, current recognition rates definitely do not allow for automated indexing of chemical document collections [92]. This is even more serious because the growing number of publication platforms, like open access journals or the demands of retro digitalization, calls for an automatic yet accurate way of indexing at least the documents' important chemical structures.

Actually, the problem of uniquely naming chemical structures in texts is not very new. For a long time, chemists have developed different algorithms for converting a chemical structure to unique line notations. Such a notation is, e.g., the *IUPAC name* which yields into a unique representation for small molecules (introduced around 1920). But for more complex molecules, the IUPAC rules are still ambiguous. Moreover, for the use in digital systems chemical names have been transformed into linear notations. Today, the prevalent linear notations are the *International Chemical Identifier* (InChI) and the *simplified molecular input line entry specification* (SMILES) which indeed are unique representations, but show high complexity and are almost impossible to dissect for humans. Therefore, they are not widely used in chemical documents and thus cannot be extracted for indexing purposes.

In fact, beside graphical representations, chemical documents refer to entities usually using trivial names and rely on the reader to figure out the contextual information. But also this does not help indexing: each chemical structure may have several different trivial names, often chosen with respect to the paper's context, e.g., pharmaceutical names, brand names, or terms from natural product chemistry. As always, the challenge for search engines using the entity name is to discover all related synonyms and disambiguate terms based on the document context. In particular, failing to index all entities may lead to the exclusion of highly relevant documents.

Facing these problems, chemical information service provider offer specialized indexes. These indexes are built up by *manually* identifying and indexing all chemical structures from a document collection in structure databases. The resulting structure databases then are accessed through graphical interfaces. By drawing a chemical structure a domain expert can thus formulate a query, which in turn will be parsed by the chemical query parser and matched against entities' fingerprints stored inside the structure database. The amount of manual work required for building up and maintaining such indexes results in high costs. Today, the most important provider is the Chemical Abstract Service (CAS[30]) offering high quality data at a price of about 30,000 USD/year for a single user subscription. Obviously for the growing open access movement this type of indexing documents is not a viable option.

[30] http://www.cas.org/expertise/cascontent/index.html

Our aim is to make the large body of chemical knowledge stored in the Web widely searchable and accessible, however, with a minimal amount of manual indexing. Therefore, our system automatically extracts chemical entities from document collections, indexes them with synonym mark-up and disambiguation, and finally makes the documents searchable by commonly used Web search interfaces, like for instance Google or Yahoo!. Hence, our contribution is twofold:

- Firstly, we developed an information service that automatically generates enhanced metadata representations from chemical documents. These metadata enrichments include extensive information for each entity found in the full-texts, e.g., trivial names with synonyms, InChI codes, SMILES, and basic chemical properties. By generating respective HTML pages and linking to the respective document sources, current crawlers can easily index the information in connection with each document. Our experiments clearly show the added value for chemical document retrieval.
- Secondly, by providing rich and diverse metadata our system is able to support typical, and even sophisticated chemical workflows. In contrast, previous approaches in digital libraries, like e.g., indexing entities by simple chemical formulae, see e.g., [61] are entirely useless from a chemist's point of view due to the ambiguities: for instance for the simple formula C_6H_6 there are already more than 200 different structures, each of them with different chemical properties and uses.

4.4.2. Use Case

Assume our scientist is interested in the synthesis of odorous substances, e.g., as ingredients for perfumes. In particular, our chemist may be looking for building blocks usable in various synthetic pathways. Here, a simple precursor is the molecule *methoxybenzene* (see Fig. 21 left), which is a common intermediate in the production of pharmaceuticals or odorous substances. In fact, a derivate of *methoxybenzene*, *1-methoxy-4-(1-propenyl)-benzene*, is the main component of anise oil (see Fig. 21 right) which can be isolated by steam distillation from star anise (*Illicium verum*) or anise (*Pimpinella anisum*).

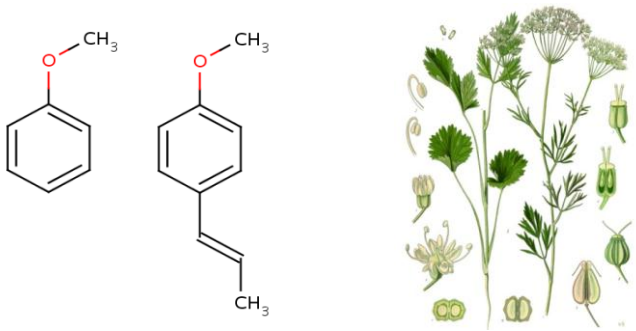


Fig. 21. Methoxybenzene and 1-methoxy-4-(1-propenyl)benzene (left)
Anise, from Koehler's Medicinal-Plants 1887 (right)

For the sake of open access assume that in his/her search for information our practitioner faces lacks access to commercially available chemical structure databases (due to the high prices or license limitations). Focusing on a name-based search our practitioner has to face the challenge of disambiguating chemical names (IUPAC, INN, trivial or brand name). Picking up our example entity *methoxybenzene*, one could also search for *phenoxymethane*, *phenyl methyl ether*, or even the trivial name *anisole*. All these names represent a valid verbal description of the substance. Therefore, our chemist first tries a keyword-based Web search using the query term '*methoxybenzene*', specifically on information from freely available open access journals.

For example, the journal *Archive for Organic Chemistry* (ARKIVOC)[31] Journal is one of the most renowned open access sources for organic chemistry, published since 2000, containing detailed experimental information about various compounds. But for the ARKIVOC collection a search for '*methoxybenzene*' returns zero hits. Still, only given the full texts it is impossible to distinguish whether the document collection simply does not contain any document with the entity or if our practitioner has only selected a verbal descriptor of the compound not used within the documents. In fact, a query on '*anisole*' would have retrieved 7 correct results. Thus, providing and maintaining a proper index, linking all relevant information about substances to the papers they occur in, is vital.

4.4.3. Experiment

For our evaluation we used a collection of 2588 chemical documents from ARKIVOC. This document collection has been processed by our system described in the last subsection resulting in a set of enriched index pages. To assess the difference between a Web search over our semantically enriched index pages and plain full-text retrieval we used a simple Lucene whitespace analyzer to build an inverted index for the full-text documents (baseline) and the enriched index pages. For structure search the chemical entities are stored in a MySQL database in a structure table constructed by ChemAxon[32]. Basically we performed four different experiments:

1. First, we evaluated the impact of our enriched index pages in terms of average result set relevance. The results of randomly chosen text queries were evaluated in a precision/recall analysis.
2. To evaluate the quality in terms of ambiguity resolution we compared the retrieval results using enriched index pages to an exact structure search.

[31] www.arkat-usa.org

[32] www.chemaxon.com

3. To show the practical applicability of our approach especially over large document collections we also compared the respective retrieval times of structure and text search.
4. Since our global aim is to expose chemical document collections hidden in digital libraries via commonly used Web search interfaces, like e.g., provided by Google or Yahoo!, we made our enriched index pages available online. Then we analyzed the number of pages crawled by Google and to what degree our pages are actually indexed.

Impact of Semantic Enrichment

In this experiment we evaluate the impact of our enriched index pages using a precision/recall analysis. Relevance can only be assessed manually by domain experts (in particular chemists), in what is a very expensive process. Therefore, we performed the precision/recall analysis only on a subset of documents (still about 10% of the entire collection). To choose a *representative* subset, we analyzed the number of occurrences of individual chemical entities in the document collection. Fig. 22 shows the distribution of the 5000 most often occurring chemical entities.

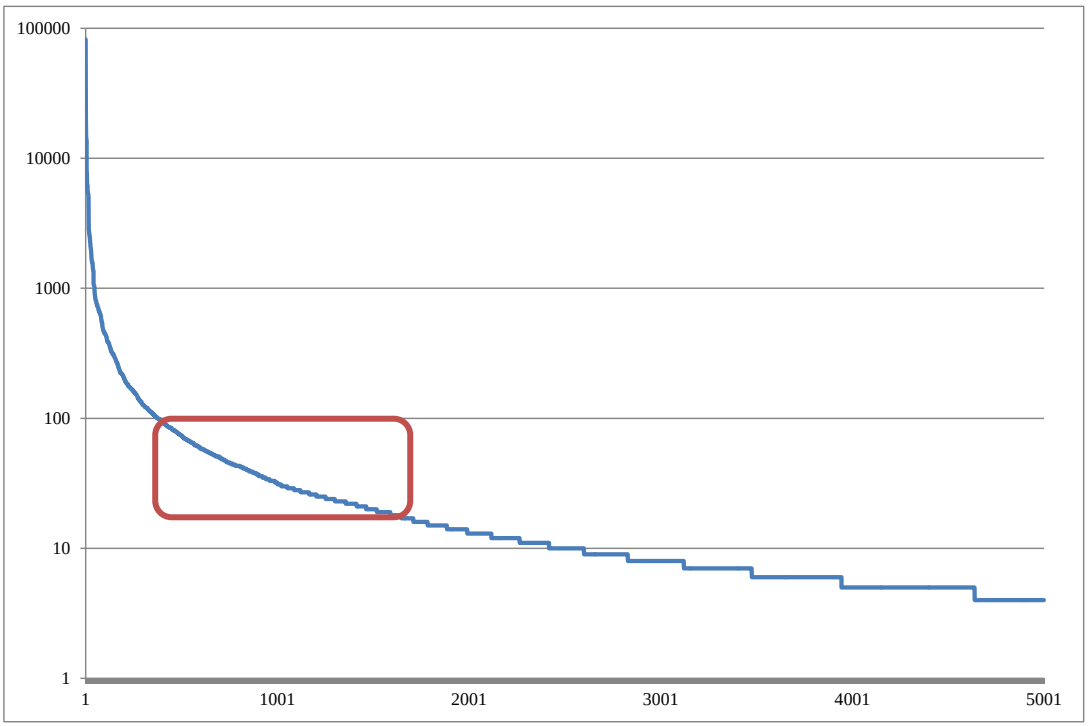


Fig. 22. Distribution of entity occurrence in documents

Since it is not sensible to choose entities for evaluation that occur either in almost all documents or are extremely rare, we chose our query entities for evaluation only from entities occurring in less than 100, but more than 20 documents. We retrieved all documents matching the queries and randomly chose a subset of

10%. From these documents we randomly selected a total of 5% of the occurring entities resulting in 22 textual query terms varying from trivial entity names to InChI codes. For the evaluation domain experts in the field of chemistry considered all retrieved documents with respect to each query and judged the relevance in a binary fashion.

To determine the practical value of our textual indexing, the domain experts used a very strict relevance rating: documents are only marked as relevant, if there was an exact match for the query entity regarding both syntax *and* semantics. For example, the relevance judgment distinguished between actual substances and substance classes. Since classes are often simply given in the plural form of the respective substance this poses a difficulty for stemming in text search engines. Even worse, in some documents complex entities are described using a basic entity name as placeholder for a more detailed structure shown in some image. Since the actual structure may have totally different chemical properties also such documents have been considered as errors in the relevance analysis. Finally, sometimes an entity name can even be used as a placeholder for describing certain characteristics or functionality of other entities, i.e. although some entity name may occur in a paper, the actual entity may not be relevant. The experts also counted such documents as false retrievals in the text search.

In total from all documents retrieved as query results the domain experts marked 158 documents as relevant regarding the respective queries. Table 5 shows the resulting precision/recall values.

Table 5. Precision and relative recall values for baseline and enriched search

Search type	Retrieved	Retrieved + Relevant	Recall	Precision
Baseline	87	58	0.37	0.67
Enriched	259	150	0.95	0.58

As expected, we experienced a very low recall value of only 37% for the baseline approach. In contrast, the recall for our enriched index pages is 95%. The semantic enrichment thus yields essential benefits. For example, there will almost never be a hit in the baseline full text documents for queries on InChI codes, whereas our index pages include the InChIs and all synonyms of the query term for most of the structures. But given the strict relevance voting necessary for practical usefulness this tremendous recall benefits have to be paid for in terms of precision. Still, the precision of our approach has only slightly decreased to 58% compared to 67% for the baseline documents. Basically, due to our enrichments the result set size grows, however, this increases also the number of technically correctly found, but semantically irrelevant documents.

Table 6. F_x-Measure values for baseline and enriched search

	F_1-Measure	**F_2-Measure**	**$F_{0.5}$-Measure**
Baseline	0.47	0.40	0.57
Enriched	0.72	0.84	0.63

To also quantify the overall benefit of our enrichment technique we computed the weighted F-Measures. Table 6 shows the different F-Measure values of the different search types. For the classic F_1-Measure we can already see a dramatic improvement of more than 0.2 over the baseline. Moreover, document retrieval in the area of chemistry is rather recall oriented: it is fatal to miss a single document related to the query. For an industrial research team missing relevant research results (e.g., with respect to patents) may lead to enormous costs for the respective company. Hence, the actually most significant measure for our scenario is the F_2-Measure weighing recall higher than precision. Here, our algorithm even scores an improvement of more than 0.4. But even when a user focuses on a precision-oriented search, our algorithm still results in a small benefit of 0.06 for the $F_{0.5}$-Measure.

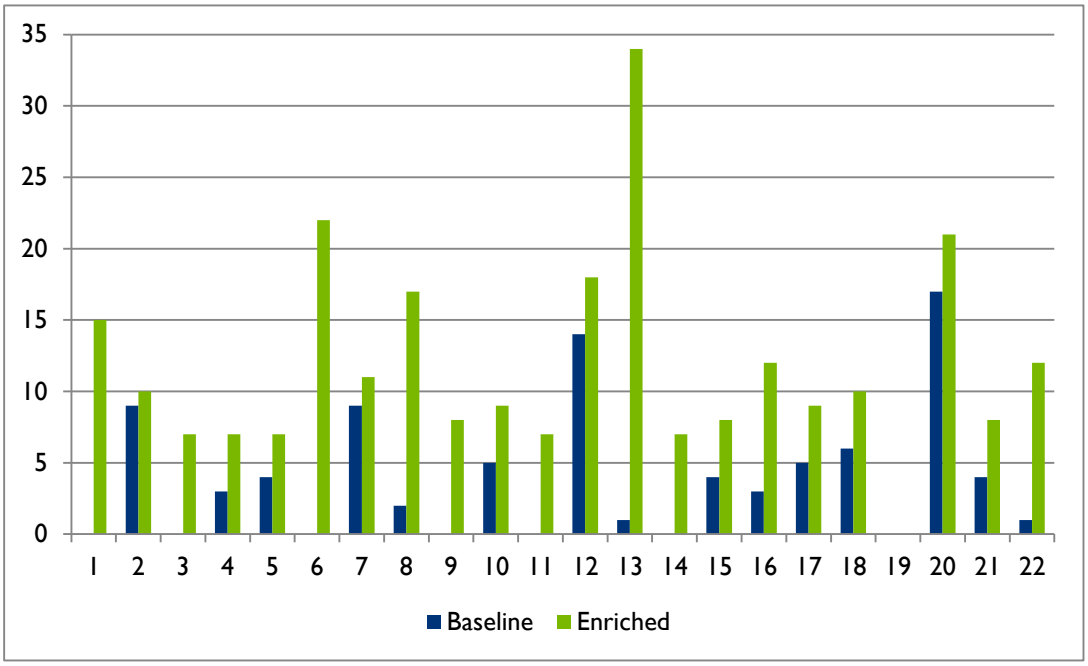


Fig. 23. Retrieved documents per query: enriched versus baseline search

Investigating the search results per query more closely we found that the benefit can really be seen in all searches. Fig. 23 shows a detailed overview of the number of retrieved documents per query. For all queries the enriched index pages retrieved more relevant documents than the baseline search. An exception is query 18 where no matching document was found in either approach. The respective

query term *InChI=1S/C5H8O/c1-2-4-6-5-3-1/h2,4H,1,3,5H2* cannot be found because the responsible entity in the original document could not be matched uniquely to the PubChem entities. As we can see, there is still need for further improvement for metadata enrichment.

Quality of Semantic Enrichment

To measure the quality of our enriched search approach we compared the results to a chemical structure search, which currently is state of the art for chemical digital libraries. But a structure search has complex requirements: it is necessary to use specialized commercial software, e.g., ChemAxon's JChem suite, to build up a structure database. The structural data is stored in a proprietary format (varying dependent on the vendor) and also the access to the data is only possible by using appropriate graphical query interfaces where structures can be sketched.

Structure search applications offer different query types: beside an exact structure search also sub-/super-structure and similarity searches are possible. Unfortunately, these search types are not directly portable to textual searches, because e.g., substructures of an entity are not simply substrings of the entity name. Therefore, we have to focus on exact matching structures in our experiments, and leave other kinds of structure searches to future work. For each of our query terms we took the corresponding structure information of the chemical entity and retrieved all matching documents. The document and query set is the same used in the previous experiment.

Table 7 shows that the relative recall value for our enriched index pages of 95% is very similar to the respective value for the structure search. And also the precision values of 58 % for enriched and 59% for structure search are almost identical. Hence, also the F-Measures shown in Table 8 are nearly the same. Please note, that although structure search has more complex requirements, it offers only a slight advantage for exact matching queries over searching our enriched index pages.

Table 7. Precision and relative recall values for enriched and structure search

Search type	Retrieved	Retrieved + Relevant	Recall	Precision
Enriched	259	150	0.95	0.58
Structure	262	154	0.98	0.59

Table 8. F_x-Measure values for enriched and structure search

	F_1-Measure	F_2-Measure	$F_{0.5}$-Measure
Enriched	0.72	0.84	0.63
Structure	0.73	0.86	0.64

Again we investigated this effect on query level. Fig. 24 compares the retrieved documents for each query entity. As expected from the precision/recall analysis, in most queries enriched and structure search retrieved the same number of documents.

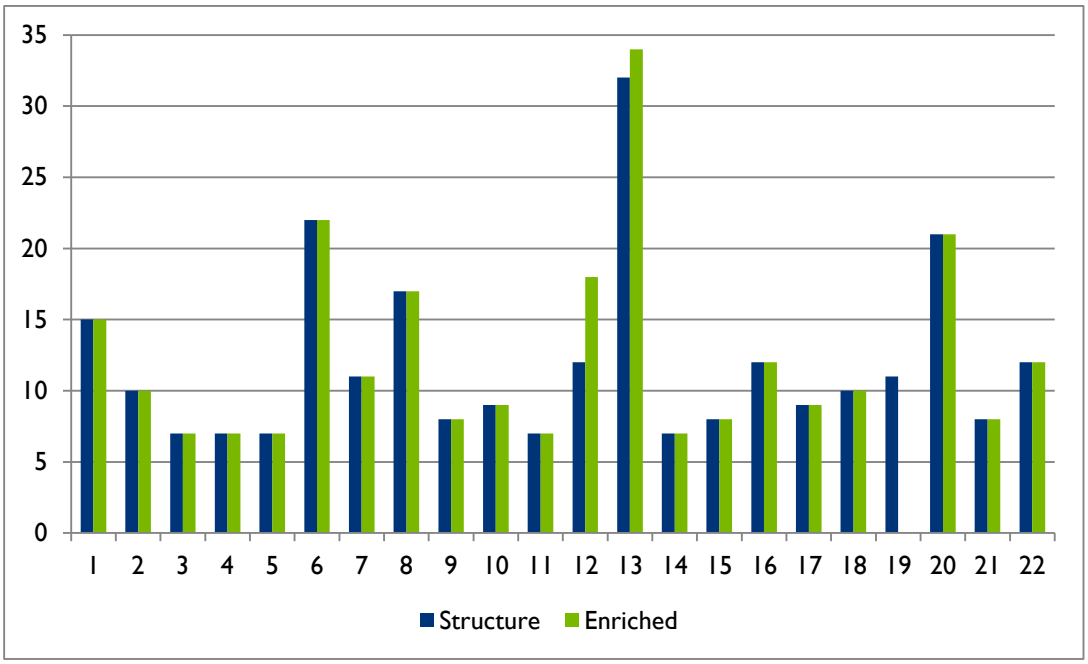


Fig. 24. Retrieved documents per query: enriched versus structure search

The only exceptions occur for queries 12, 13 and 19. We already commented on the ambiguous entity term in query 19; of course a structure search can resolve this ambiguity accounting for the slightly increased recall of structure search. Moreover, for queries 12 and 13 some irrelevant documents were found in the text search, because the query entity was a substring of some more complex entity occurring in the document. For example, the query term for query 12 is *iodobenzene*. Here, also irrelevant documents containing entities like e.g. *diacetoxyiodobenzene* or *tetraiodobenzene* are retrieved. Also the abbreviated naming of entities by using their functional groups only contributes to the false retrievals.

To summarize this experiment, we can state that a text search on enriched index pages indeed yields similar results to a chemical exact structure search with respect to the retrieved documents.

Search Performance

In this experiment we compare the respective retrieval performance in terms of response times for text- and structure search. The measured time comprises query processing until all relevant documents have been retrieved. We performed experiments over several days on our digital library server to get representative average values. We did three batches, each run including 10.000 queries, varying

the query terms for the text search between SMILES, names and InChI codes. The 10.000 query entities were chosen randomly from our entity database. For the structure search always the SMILES code is used which is internally converted into a unique structure representation of the respective entity. Please note, that usually also the drawing of the actual structure followed by a conversion into a SMILES code or CML would be part of the structure search. We discounted these costs by directly starting from the SMILES code. In any case, the conversion of linear notations to fingerprints is a step that has always to be performed in structure search independently of whether a SMILES code is directly given or the structure is drawn. After finding the exact matching entity for that structure all related documents are retrieved.

In text searches beside single term queries also query terms concatenated with Boolean operators are commonly used. Therefore, we simulated 'AND' and 'OR' searches. Since in structure search Boolean queries are not easy to perform, the only way here is to make two subsequent structure searches.

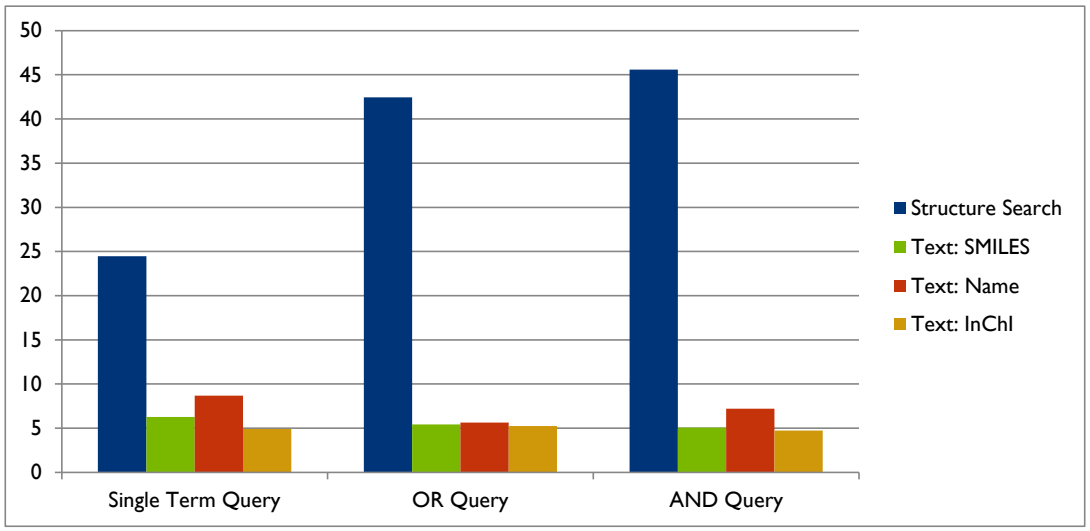


Fig. 25. Retrieval times [ms] for different search types

Fig. 25 shows the average retrieval times measured for the different search types. As a general trend we can see that text searches are far more efficient. For instance, in text search it makes no difference which query term is used or if more than one term is concatenated in a Boolean query. The retrieval times only vary between five and eight milliseconds (note that name search is slightly less efficient than SMILES or InChI, because of many synonyms). Using a structure search the document retrieval is always about an order of magnitude slower due to the complex matching of fingerprints. Moreover, the time for queries using Boolean operators is rather high, since here two (or more) structure searches are needed (in our experiments we only used simple queries comprising two terms).

In summary, our results show that a text search is always much faster than a structure search independently of the text search's query term. Moreover, for Boolean queries the retrieval time for text queries does not increase.

Indexing for Web Search

Our overall aim is to improve access to chemical document collections hidden in digital libraries via common Web search providers. Therefore, we simply made all enriched index pages for the ARKIVOC journal available on the Web. To have a chance of being indexed the generation and layout of our enriched pages is important. Most crawlers would mark pages within a site as spam, if they just show some index terms and do not include at least some full text or links. Therefore, our pages include, beside the actual enriched metadata table, the document's title, its abstract and a link to the full text. On the other hand, high quality open access journal will also feature high PageRank, thus crawlers will index them prominently.

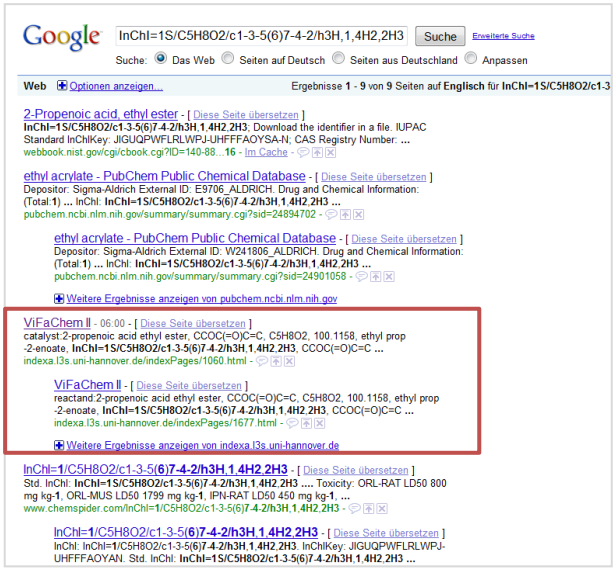


Fig. 26. Google Search Example for InChI code

After three month of being online the Google index indeed contained already around 600 of our pages. However, it is not traceable how the pages are indexed and exactly why a page is indexed and some other not. Fig. 26 shows a screenshot of a text search on the term '*InChI=1S/C5H8O2/c1-3-5(6)7-4-2/h3H,1,4H2,2H3*'. The enriched index page for the relevant ARKIVOC journal paper '*Effect of substituents and benzyne generating bases on the orientation to and reactivity of haloarynes*' appears on third place in the Google result, directly after the respective dictionary entries of the substance from the National Institute of Standards and Technology and the PubChem substance database.

Although we did nothing to promote the index pages, i.e. our pages still have a Google PageRank of 0 (as opposed to PageRank 7 for both NIST and PubChem),

they are still found and provide access to relevant documents that would not have been found otherwise (as the respective ARKIVOC journal papers do never appear in the Google search result). Please note that for investigating the indexing process we always chose '*ViFaChem II*' as title for all of our enriched pages to detect them easily in the Web search results. Of course, usually the journal name and title of the related document is used.

4.4.4. Conclusion

Measuring the quality of an index is not limited to the classical measures, i.e. correct retrieval of documents and interindexer consistency anymore. By using semantic techniques, it is also essential to consider the quality of these techniques (see section 4.3), because they are used to enrich the indexes by, e.g. synonyms. Thus the interindexer consistency may vary depending on the used semantic technique. Indeed this does not mean that the quality of the index itself is corrupted.

Today, an additional pitfall is the user behavior itself: A keyword based Web search is the standard for the information seeking process for most of the users. Thus, most digital library indexes are *hidden* for these users, because they are not exposed to Web search engines. The idea of our approach is to open up literature hidden in digital libraries by simply enabling text queries in commonly used search interfaces, like e.g., provided by Google or Yahoo!.

We have shown one solution in the domain of chemistry. To facilitate this, we had to solve several problems. Chemical substances can have many different and often ambiguous textual representations, like trivial names, InChI codes or SMILES. In chemical documents besides structure images usually only trivial names are used for brevity and improved readability. We developed a workflow allowing the automatic generation of customized index pages including all metadata information extracted from publicly accessible databases for each occurring chemical entity. Our framework can easily be used, e.g., by libraries, open access journals, or other content providers in the chemical domain. We also performed experiments to show the usefulness of our approach. The retrieval quality of our enriched index pages is almost as good as chemical exact structure searches and significantly better compared to a baseline/full text search.

Based on the results of our experiments, it can be stated that it is possible to open up hidden document collections to Web search engines in such a way, that domain specifics are considered and the index quality does not suffer.

4.5. Document Retrieval

In the last section we discussed the problem of building a semantic enriched index in such a way, that also hidden collections are available with a simple Web search. Today, it is very important to open up the digital library indexes, because a keyword based Web search is the starting point for almost all information gathering

processes. However, in some highly specialized domains a simple keyword based search is not sufficient. For example, the information gathering process in not only biology, medicine and chemistry is entity centered.

In the last section we have shown how semantic metadata can be used for building up index pages for documents. These index pages are indexed by Web search engines, e.g. Google and linked to the original documents. Since synonyms and different entity representations are included in the index page a Boolean string based entity search retrieves almost all desired documents. However, domain experts are usually searching for a very specific entity occurring in a specific task. Besides the entity this task is also necessary as query term to enhance the precision. Therefore, the search process is based on a combined Boolean query including both the entity and a specific task. For example, a biologist could search for a specific gene in the context of *human* or *mouse,* whereas a medic can search for a specific drug in the context of a specific side effect of this drug. In most cases the combination of entity and specified task occur very seldom, resulting in a low number of retrieved documents (overspecialization). The task formulated as second query term imposes a hard constraint and cannot be relaxed. Considering e.g. the chemical entity, there are several other substances having the same functional properties. Therefore, it is inevitable to relax the first query term and search for entities with the same or similar properties.

In general, in the context of personalized document retrieval in a digital library we always have to face the problem of query overspecialization. A well-known approach to address this problem is query relaxation, see e.g. [93–95]. But therefor, it is in general essential to compute *similarities* between entities. However, it is not trivial to compute similarities between entities, because the measures for this purpose differ a lot according to domain and task. Consequently, for high quality document retrieval, we have to assess the quality of the used similarity measures and the resulting retrieval process within our life cycle.

In this section we build an exemplary personalized retrieval system to overcome the problem of overspecialization in the context of chemical digital libraries. For computing similarity between chemical entities, the first necessary step is to convert the entities to a fingerprint representation. There are numerous fingerprints available, all of them emphasizing different attributes of a chemical entity, e.g. structural information, functional groups or number of atoms. Beside the different fingerprint representations, more than 40 similarity measures for chemical entities are available. In order to better understand the background story we first examined to which degree the similarity measures are correlated. The uncorrelated measures are further used in a feedback step in our system to learn which measure is most appreciated by the individual user. We evaluate which combination of fingerprint and similarity measure is useful for personalized query relaxation.

4.5.1. Use Case

The following scenario is typical for the daily work of a practitioner in the chemical domain. Imagine a chemist from the field of drug design who is currently working on an improvement of Viagra®. In this scenario he is searching for related literature about the active ingredient, *Sildenafil (see* Fig. 27*)*.

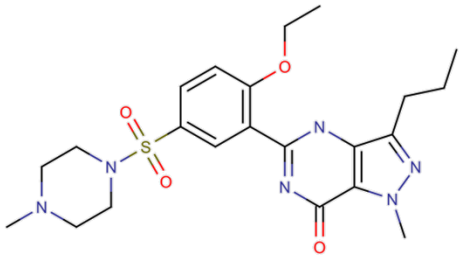


Fig. 27. Structure of Sildenafil

Usually, the access to chemical literature is performed by a drawn query, using a specialized search interface. Today's chemical search engines are also able to relax the query entity by searching for entities with corresponding substructures or entities with similar properties. Problematically, all documents including these entities are returned. To overcome the obstacle of a specialized interface, we introduced an approach to open up the chemical domain for text based search engines [96]. The system generated an index page, containing all chemical entities included in the respective document, for each document in the digital library. Beside the entity name found in the document also all synonyms and different representations, like e.g. SMILES or InChI code are included. The evaluation has shown that the results were almost as good as a chemical structure search.

Chemistry is a wide field and chemical entities are usually used in many different contexts, e.g. drug design. Our chemist wants to overcome one specific side effect of *Sildenafil*, namely '*irregular heartbeat*' he is searching for documents describing this side effect. Since, to the best of our knowledge, current chemical search engines do not support the search for entities occurring only in a specific context, our chemist has to manually scan all retrieved documents.

Certainly, it is possible to build a search engine which has the ability to combine the entity and context as a query term. A simple architecture dealing with these combined queries is shown in Fig. 28. Here, the combined user query *Q!* is sent to the search engine and split up into the chemical entity E_q and the specified context q_i. The documents including q_i can easily be computed using an inverted full text index. Searching for relevant documents regarding E_q is more difficult since we have to take all different entity representations (e.g. SMILES or InChI codes) and synonyms into account. To address ambiguity, we rely on chemical index pages (see [96]) to search for relevant documents. The intersection of both result sets shapes the final result set and is delivered to the user. Note that, due to the fact

that in chemical documents the most relevant entity, i.e. the product of a synthesis, can occur only once, only Boolean queries are reasonable and traditional IR measure, e.g. TF*IDF, are not.

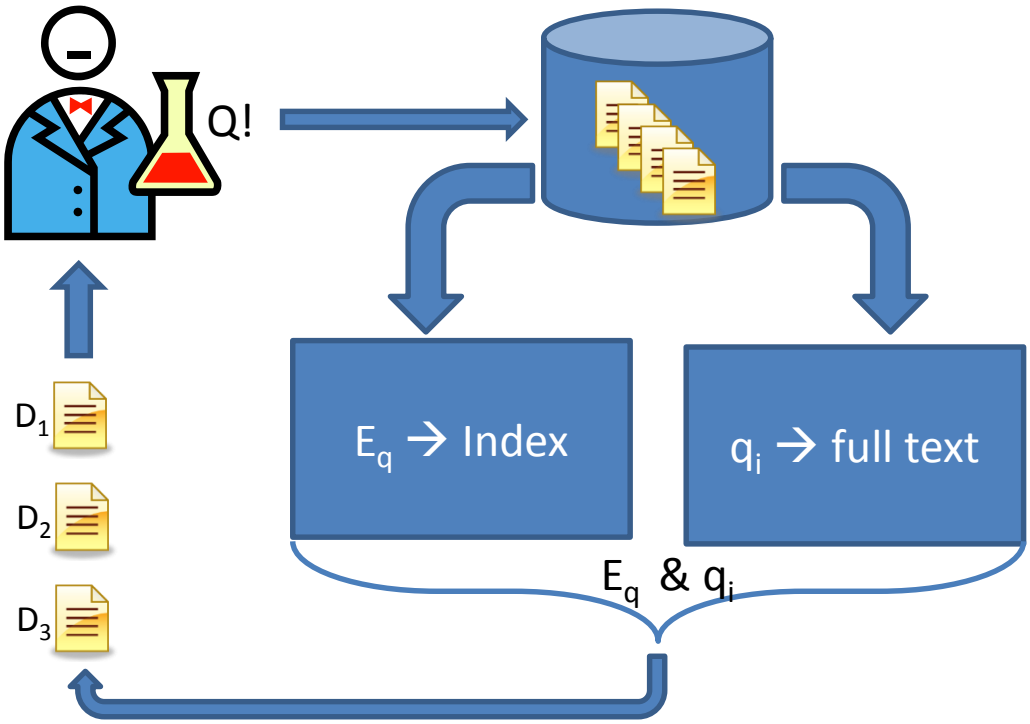


Fig. 28. Simple architecture

Focusing on our chemist again, he can now search for documents about '*Sildenafil*' and '*irregular heartbeat*'. Unfortunately, he is still unable to fulfill his information need, because the active pharmaceutical ingredient '*Sildenafil*' is trademarked and cannot be used for other drugs. As a consequence, he must relax his query to find other chemical entities with similar properties within the same context. Indeed, this query relaxation should be done automatically by replacing the actual entity with similar entities. In the following sections we will extend the simple architecture to develop such a system.

4.5.2. Fingerprints and Similarity Measures

For similarity computation between chemical entities many different measures are available. As we will see later, some of them are uncorrelated because their result sets differ. The first necessary step for computing similarity is the transformation of a chemical substance into a fingerprint. Since a lot of fingerprint transformations are available, the amount of possible combinations of fingerprints and similarity computations between them is really high. The idea of measuring the similarity of two objects, each defined by a set of common attributes, is discussed in many different domains, including e.g. biology [97] or chemistry [98]. Although these appli-

cation areas are divers, the used similarity coefficients are almost the same. Since the performance always relies on the choice of an appropriate measure, many researchers have worked on finding the most meaningful measure. The work done by Willet et.al, [98] and [99] gives overviews of the coefficients that have found widespread use in chemical information systems.

Even though numerous binary similarity measures have been described in the literature by their properties and features [100–103], only a few comparative studies are available. In the field of biology Hubalek collected 43 similarity measures and after evaluating similarities, correlations, transformations of the value range and symmetry, 23 were excluded. The remaining ones were used for cluster analysis on fungi data to produce five clusters of related coefficients [97]. In the domain of chemistry, Willet evaluated 13 similarity measures for binary fingerprint code [104]. Current work identified the most useful fingerprint based chemical similarity measures [99]. We use these measures and combine them with different fingerprint representations to identify correlation between them.

Fingerprints encode molecular structures in a series of binary digits (bits) where bits are set according to occurrences of particular structural features. For generating fingerprints, the structure is converted into its unique SMILES representation [105]. There are several ways of creating fingerprints focusing on different fragments of chemical entities. Examples for typical fragments for generating fingerprints are:

— *Atom sequence*: A linear path of atoms and bonds through the molecule.
— *Ring composition*: An atom and bond sequence around a ring structure in the molecule.
— *Atom pairs*: A pair of atoms in the same molecule with number of bonds in the shortest path between them. The different atom pairs are usually further differentiated by, e.g., taking the number of attached hydrogens into account.

Sometimes fragments are too specific, leading to very low frequencies and sparse fingerprints, included atom and bond types can be generalized. We rely on the open source chemical development toolkit (CDK) [106] which includes the following fingerprints.

Standard Fingerprint This fingerprint examines the molecule and encodes the following:

— a pattern for each atom
— a pattern representing each atom and its nearest neighbors
— a pattern representing each group of atoms and bonds connected by paths up to 2 bonds long
— a pattern representing the atoms and bonds connected by paths up to 3 bonds long
— a pattern representing the atoms and bonds connected by paths up to 4, 5, 6, and 7 bonds long

Extended Fingerprint An Extended fingerprint includes in addition to the Standard fingerprint features for describing aromatic rings.

Graphonly Fingerprint This fingerprint is a specialized version of the Standard fingerprint that does not take the bond order into account.

EState fingerprint generates 79 bit fingerprints using fragments describing the electronic and topological characterization of an atom, called electrotopological state (e-state) [107]. The fingerprint simply indicates if such a fragment is present in the structure or not.

Substructure Fingerprint currently supports 307 different substructures. A set bit indicates that the related substructure was found in the molecule.

MACCS Fingerprint is the representation of the answer of 166 questions about a chemical structure [108].

Considering these fingerprints, we examined the most common useful measures (see Table 9) in the domain of chemistry collected in [99]. The variables of the formulas are defined as follows: If we consider two fingerprints of two chemical entities A and B, then:

— *a* is the count of bits set to 1 in entity A but not in entity B
— *b* is the count of bits set to 1 in entity B but not in entity A
— c is the count of the bits set to 1 in both entity A and entity B
— *d* is the count of the bits set to 0 in both entity A and entity B

Table 9. Reviewed similarity measures

Measure	Range	Formula
Cosine	[0, 1]	$\frac{c}{\sqrt{(a+c)*(b+c)}}$
Dice	[0, 1]	$\frac{2*c}{(a+c)*(b+c)}$
Euclidean	[0, 1]	$\sqrt{\frac{c+d}{a+b+c+d}}$
Forbes	[0, ∞]	$\frac{c*(a+b+c+d)}{(a+c)*(b+c)}$
Hamman	[-1, 1]	$\frac{(c+d)-(a+b)}{a+b+c+d}$
Jaccard / Tanimoto	[0, 1]	$\frac{c}{a+b+c}$
Kulczynski	[0, 1]	$0.5*\left(\frac{c}{a+c}+\frac{c}{b+c}\right)$
Manhattan	[1, 0]	$\frac{a+b}{a+b+c+d}$

Matching	[0, 1]	$\frac{c+d}{a+b+c+d}$
Pearson	[-1, 1]	$\frac{(c*d)-(a*b)}{\sqrt{(a+c)*(b+c)*(a+d)*(b+d)}}$
Rogers-Tanimoto	[0, 1]	$\frac{c+d}{(a+b)+(a+b+c+d)}$
Russell-Rao	[0, 1]	$\frac{c}{a+b+c+d}$
Simpson	[0, 1]	$\frac{c}{\min\ ((a+c),(b+c))}$
Tversky	[0, 1]	$\frac{c}{\alpha*a+\beta*b+c}$
Yule	[-1, 1]	$\frac{(c*d)-(a*b)}{(c*d)+(a*b)}$

Correlation Analysis

Since now, there is no work done in the literature, analyzing the correlation of the similarity measures applied on different fingerprints. Thus, our first goal was to explore if the underlying fingerprint has some influence on the similarity measures.

To do our first experiment, we took a random 1‰ sample of the PubChem database resulting in around 44.000 chemical entities. We downloaded their SDF files to have the structural information of all entities and converted them into their respective SMILES representations. These SMILES codes were necessary to generate the different fingerprint representations of each chemical entity using the CDK. In addition, we randomly choose 20 chemical entities as query entities. Since, in a later step, we want to use the similarity measures for a personalized retrieval system it seems reasonable to evaluate not only the complete result set of around 44000 entities but also smaller subsets. Thus, we decided to also evaluate the differences between the top-x results. Therefore, we computed for each combination of fingerprint, chemical entity and top-x the 16 fingerprint based similarity measures resulting in around 88 million similarity values.

As we can interpret the similarity value as a value in a ranking vector, we decided to use the Kendall rank correlation coefficient (KTau) [109] to determine the correlation of the different measures and fingerprints. We calculated the correlation coefficient for each ranking vector and the arithmetic mean over 20 queries. A KTau of 1 means that the agreement of two rankings is perfect, -1 indicates a perfect disagreement and for independent rankings one would expect the coefficient to be *approximately* 0. Our experimental results have shown that the actual KTau values strongly differ over the fingerprints. For example the KTau value for the combination ‘Euclidean / Russell-Rao / EState fingerprint’ and ‘Euclidean / Russell-Rao / Standard fingerprint’ varies from 0.53 to -0.30 (see Table 10).

Table 10. Similarity measures with highest variances over EState (1), Extended (2), Standard (3), Graphonly (4), MACCSS (5) and Substructure (6) fingerprint

Similarity Measure	1	2	3	4	5	6
Tanimoto / Euclidean	0,83	0,12	0,11	0,39	0,67	0,76
Cosine / Matching	0,82	0,05	0,04	0,40	0,67	0,76
Dice / Rogers Tanimoto	0,83	0,12	0,11	0,39	0,67	0,76
Euclidean / Russell-Rao	0,53	-0,29	-0,30	-0,09	0,38	0,33
Manhattan / Russell-Rao	-0,53	0,29	0,30	0,09	-0,38	-0,33
Tversky / Forbes	0,48	-0,11	-0,09	0,23	0,17	0,54
Forbes / Kulczynski	0,39	-0,40	-0,35	0,14	0,04	0,41
Hamman / Russell-Rao	0,53	-0,29	-0,30	-0,10	0,37	0,32
Jaccard / Rogers Tanimoto	0,83	0,12	0,11	0,39	0,67	0,76
Kulczynski / Euclidean	0,83	0,00	0,01	0,43	0,68	0,76
Matching / Russell-Rao	0,53	-0,29	-0,30	-0,09	0,38	0,33
Pearson / Russell-Rao	0,73	0,10	0,11	0,33	0,60	0,59
Rogers Tanimoto / Russell-Rao	0,53	-0,29	-0,30	-0,09	0,38	0,33
Russell-Rao / Rogers Tanimoto	0,53	-0,29	-0,30	-0,09	0,38	0,33
Simpson / Euclidean	0,66	-0,17	-0,11	0,32	0,48	0,55
Yule / Russell-Rao	0,67	0,01	0,02	0,19	0,50	0,49

Due to the definition of the KTau, it is not straight forward to depict the uncorrelated similarity measures because *approximately zero* is not a well-defined threshold. To ensure a relatively high likelihood of correlation, we defined a threshold of 0.8. Based on this threshold, we evaluated how many uncorrelated similarity measures we have for each fingerprint. The results are shown in Fig. 29. Interestingly, the EState fingerprint always has the minimum number of uncorrelated similarity measures.

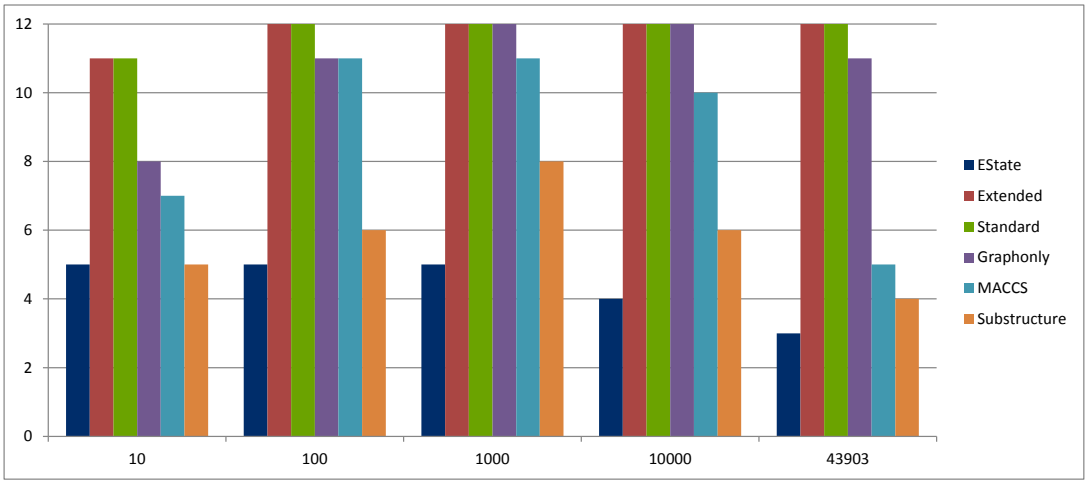


Fig. 29. Number of minimal independent rankings for top-x and a threshold of 0.8

Still, the concrete number differs from 5 to 3, which means that we have to take at least 3 different similarity measures (i.e. Yule, Russell-Rao and Forbes) into account. Given this result we notice that taking only the correlation coefficient into account is not discriminative enough; thus we consider additional discriminative properties.

Task Based Analysis

This huge variety of uncorrelated similarity measures is eligible because chemical similarity differs according to the task a chemist is working on. Intuitively, we consider that each measure might be useful for a specific task and therefore conducted experiments with example tasks using synthesis and drug design.

For drug design we took, among others, *Sildenafil* as query entity. The idea is to retrieve alternative substances with similar chemical properties. Let us consider there are two scientists from the area of drug design Peter and Bob. Both are searching for *Sildenafil*, but with different additional conditions. Peter is interested in *pyrazolopyrimidinones* with a piperazine ring system connected to the sulfonyl group. In contrast to Sildenafil Peter is looking for a free N-side at the piperazine to examine further reactions at this position. A good hit for this query scenario is *Demethylsildenafil* (see Fig. 30).

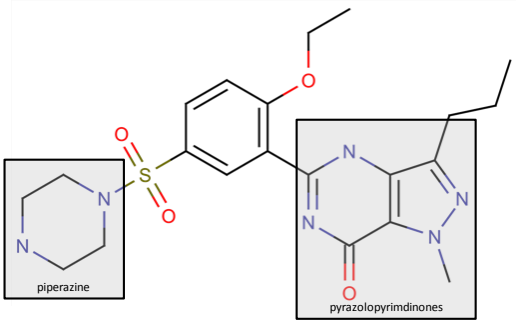


Fig. 30. Demethylsildenafil

Bob is interested in *pyrazolopyrimidinones* with a secondary amine connected to the sulfonyl group as he is interested to perform alkylation reactions at his position. *Udenalfil* with its N-alkylated secondary amine side chain represents a top candidate for this kind of query (see Fig. 31).

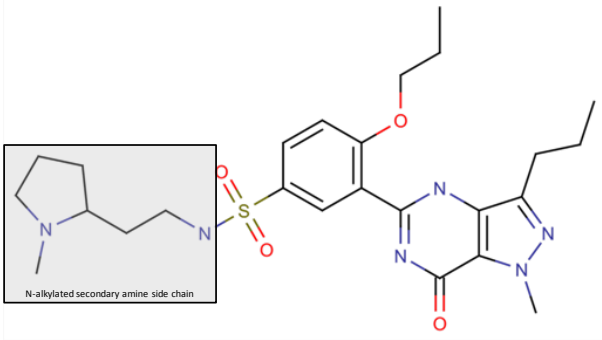


Fig. 31. Udenafil

To evaluate the ranking results of the different similarity measures, we took all chemical entities that were retrieved by a similarity search in PubChem for the query term *Sildenafil*. We also assured that the entities of interest defined by the domain experts, *Demethylsildenafil* and *Udenafil*, are included in this set. We computed similarity values for *Sildenafil* and each entity in this set using all uncorrelated similarity measures. The domain experts analyzed all result sets and evaluated which similarity measure retrieves the best result. The output of the experiment is that there is no suitable measure delivering both as relevant defined entities under the top-10. For Peter who expected *Demethylsildenafil* as relevant hit the combination of EState fingerprint and Tanimoto measure delivers the best results, ranking *Demethylsildenafil* on rank 9 and *Udenafil* on rank 335. For Bob expecting *Udenafil* as most relevant entity the combination of Substructure fingerprint and Tanimoto measure gives the best result, ranking *Udenafil* on rank 2 and *Demethylsildenafil* on rank 228. Although both chemists are from the field of drug design, they expect different ranking results for the same query term. Therefore, it is not possible to use one fixed similarity measure for one specific task. Of course, we also tried

queries for the other tasks but with the same result: it is not possible to assign one similarity measure to a specific task.

To better judge the impact of the task, we interviewed a group of domain experts to find reasons for this behavior. We figured out that each individual chemist has some kind of special background knowledge or experiences that he implies, like e.g. costs for synthesis or which substances are already in the fund of the company. This background knowledge cannot be expressed by the query term resulting in insufficient result sets.

Feedback Analysis

The task based experiment has shown that there is a need for personalized retrieval systems. The idea is to build a system where each individual user trains the system and the system will learn the similarity measure which fits best to his needs. Consequently, we conducted a user study with domain experts from the area of drug design and synthesis, to discover if already a simple feedback step would result in an explicit combination of similarity measure and fingerprint. Furthermore, we are interested in the number of feedback cycles that are necessary until such a system is stable.

For the user study, we have randomly chosen 10 query entities from PubChem, each of them representing one feedback cycle inside the system. Based on our previous results we used the 5 uncorrelated measures Russell-Rao, Yule, Forbes, Simpson and Manhattan for calculating the similarity values. In a first step, we retrieved the top-10 entities for each similarity measure and put them in one set which did not include duplicates and was unranked. In a second step, the chemists marked all relevant entities resulting in their personalized ranking vector. For each query we took the respective ranking vector and compared it to the top-10 vector of the uncorrelated similarity measures by computing precision at 10.

As an illustrating example we take the results of the domain expert introduced in our use case scenario searching for *Sildenafil* (Fig. 32). One can see that there are perfect candidates for the personalized similarity measure, i.e. a combination of the Extended fingerprint and the Yule, Forbes or Simpson measure.

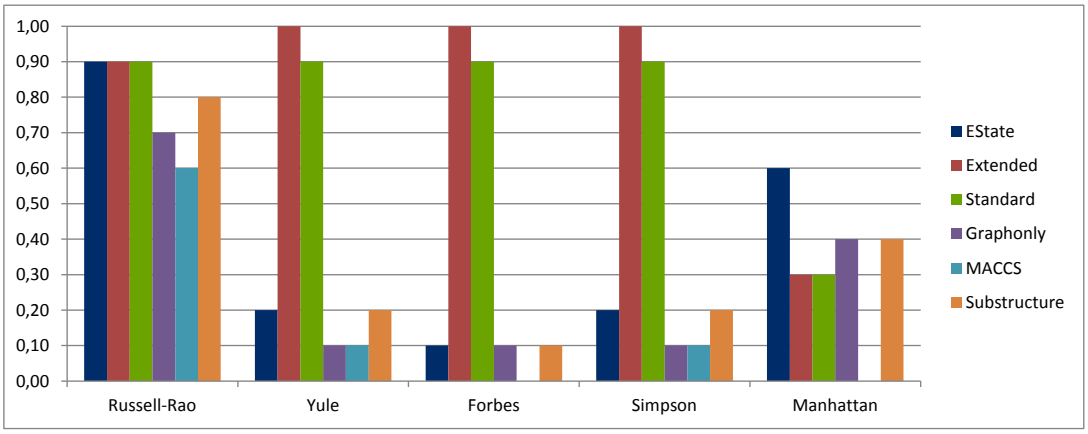


Fig. 32. P@10 values for the query *Sindenafil*

However, of course one query is not enough to decide for a specific similarity measure. Fig. 33 shows the average precision at 10 values for the chemist regarding 10 different queries. Regarding all queries the personalized similarity measure has slightly changed. Finally, the best matching similarity measure is Russell-Rao based on the Graphonly fingerprint. Only six feedback cycles were necessary to find this ideal combination for this chemist, meaning the preferred similarity measure did not change again after 6 queries. The second best measure is the combination of Yule and the Extended fingerprint.

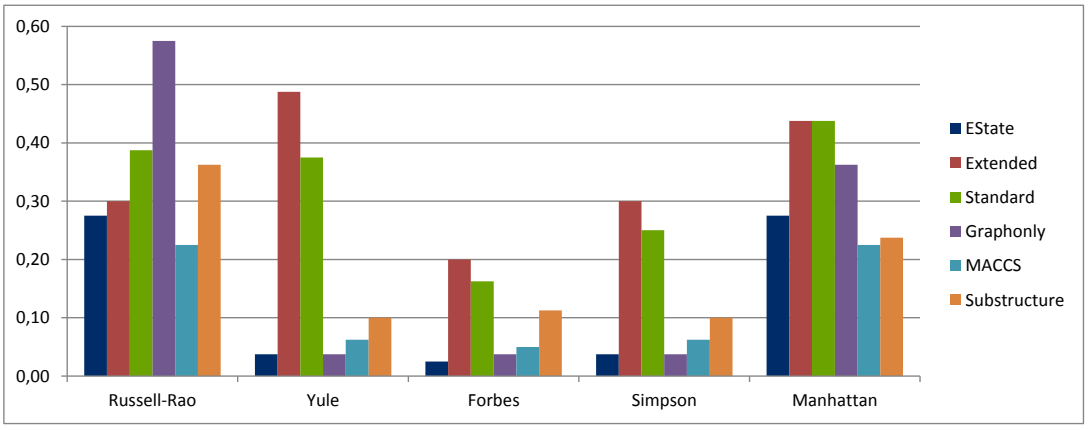


Fig. 33. Average P@10-values for one chemist over all queries

The second question to evaluate was the number of needed feedback cycles until the system was stable for an individual user. For this purpose, we defined the system as stable, if a precision value did not change more than 2% over 3 queries. We can state, that for 75% of the domain experts, the system was able to determine an explicit combination of similarity measure and fingerprint within our ten feedback cycles. The particular number of needed feedback cycles varies between 3 and 8. For the remaining 25% we could not determine a combination after 10 feedback cycles.

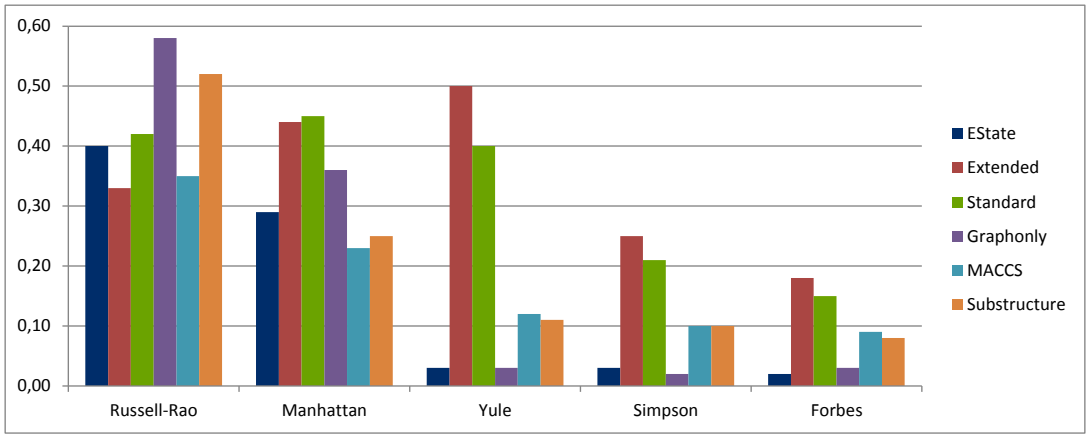


Fig. 34. P@10 values for arithmetic mean over all experts and queries

Furthermore, we analyzed the arithmetic mean over all experts and queries (see Fig. 34). One can see that the Russell-Rao measure outperforms all other measure applying it on the EState, Graphonly, MACCS and Substructure fingerprint. The best measure for the Extended fingerprint is Yule and for the Standard fingerprint it is Manhattan. Remember, these results cannot be applied out of the box to all users because the individual expectations can differ a lot. However, they are candidates for solving the well-known *new user problem*, if the user decides at least on a specific fingerprint or taking the overall best measure for a global starting point, i.e. the combination of Russell-Rao and the Graphonly fingerprint. In the next section we will describe how to integrate our findings into an architecture of a chemical search engine.

4.5.3. System Architecture with Feedback Component

We now revisit the plight of our chemist posed in the use case scenario (section 4.5.1). His aim is to find relevant documents dealing with the chemical substance E_q. Since literature for *Sildenafil* covers a lot of different topics, he further restricts the query by entering the context he is interested in, namely the side effect of *irregular heartbeat*. Fig. 35 shows our advanced architecture dealing with such kind of queries. In addition to the simple architecture, we add a query relaxation module to be able to relax the query part E_q with similar entities. The final result set only includes documents containing the context term q_i and the chemical substance E_q or q_i and at least one other similar entity for example $E_{q'}$. The document result set is ranked according to the similarity value of the included entities. As a result of the ranking function, the documents containing q_i and E_q are always top ranked followed by documents including q_i and the most similar entity $E_{q'}$.

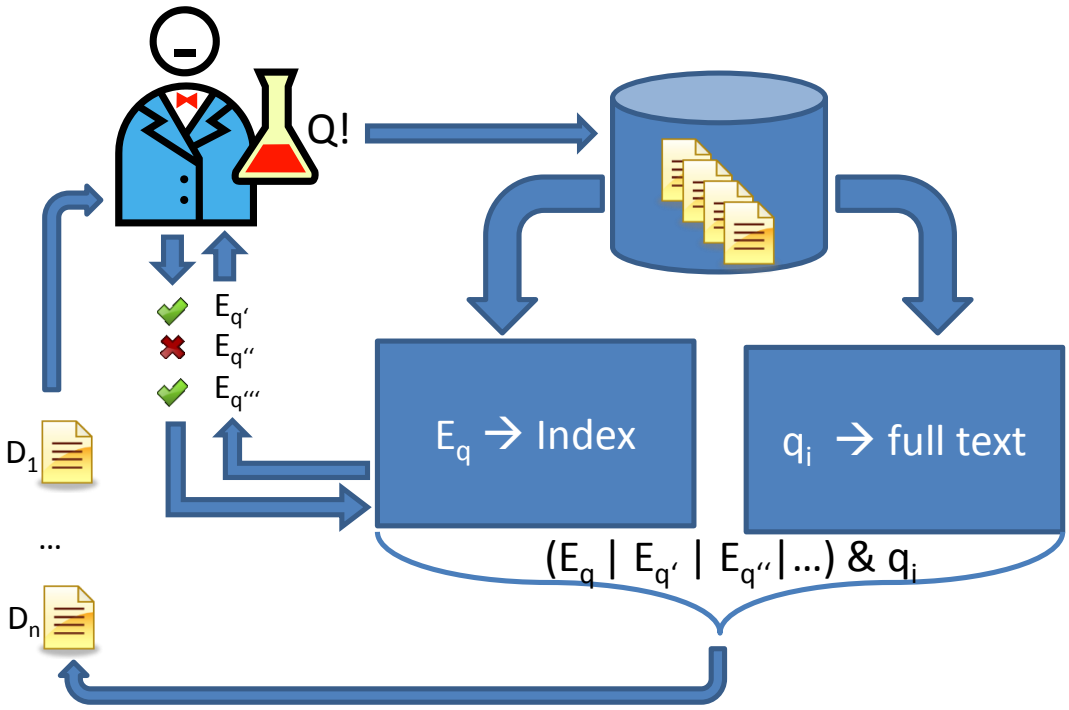


Fig. 35. Advanced architecture

As described in the previous section, a lot of uncorrelated measures are available resulting in totally different rankings and it is not obvious which similarity measure / fingerprint combination is most applicable. For a new user, the system uses the *best* similarity measure by computing the arithmetic mean over all available user feedbacks and learns the best individual similarity measure in some feedback steps.

For each feedback step, the system is calculating the top-x results of all uncorrelated measures for a query. Out of this list, the user has to decide which chemical entities are relevant for him. In a next step, the system calculates the precision at 10 values for each measure and uses the best matching one. If the chosen measure does not change over a number of different queries it is accepted as default measure for this user and the feedback step is skipped for subsequent queries. Of course, if the user is not satisfied by the proposed ranking, he can force the system to learn or to use another measure.

4.5.4. Discussion

In the context of digital libraries and especially in topic centered information portals, we have to face the problem of query overspecialization. Using the well-known approach of query relaxation, we come to the next problem, i.e. the meaning of similarity. The definition of similarity varies from domain to domain and thus it is needed to judge the usefulness of similarity measures for each of them.

As an example, we performed an experiment in the domain of chemistry. We took 16 widely used similarity measures for chemical entities and analyzed the correlation between them using Kendall's Tau. The results show that many of them are uncorrelated, meaning they deliver different rankings.

Chemistry is a wide field with many different subdomains. Therefore, chemists are focused on specific tasks when searching for literature, for example drug design or synthesis. We have analyzed whether the uncorrelated similarity measures, which are based on different fingerprint representations, fit to typical search tasks in chemistry. The different fingerprints represent different chemical aspects. For example, the Substructure fingerprint only considers the structure of a molecule, whereas the MACCS fingerprint uses a set of questions regarding more properties of a molecule than just the structure. We investigated if it is possible to assign one similarity measure to one specific task. We conducted a user study with domain experts and have shown that for the same task, e.g. drug design, different domain experts preferred different similarity measures. Hence, it is not possible to assign one similarity measure to one specific task, meaning there is no similarity measure always delivering the most suitable result set for that task. During discussions with domain experts we discovered that chemists usually have special background knowledge when searching for literature that cannot be expressed in the query, like e.g. costs for synthesis or which substances are already in the fund of the company.

These experiments have shown that it is indispensable to develop personalized retrieval systems to provide a satisfying information seeking process. We have introduced one possible solution, including a feedback cycle analyzing which similarity measures retrieve the best results for each individual user.

4.6. Conclusion

We introduced the information life cycle as a foundation for a digital library workflow (Fig. 6). We adapted the general information life cycle to be able to use the common terminology of a digital library and identified that for each step quality assurance is essential. Still, a crucial weakness of the information life cycle is that each unit of information is treated independently. Of course, individual documents and metadata records have a life of their own as represented. However the real power of networked information is the ability to follow their information and value chain. Primarily this feature brought the power to the World Wide Web and the Web Search Engines. Therefore, mechanisms should capture the provenance of data as they move through the life cycle. In our case, an ideal system should connect the original source, the used semantic technique and the resulting metadata record. In this way, we are able to reconstruct the origin of a metadata record.

Another weakness of the life cycle is the assumption that information units are fixed objects. Of course, library approaches to information management have been predicted on preserving fixed objects over time resulting in the metadata concept.

Metadata is intended to provide contextual information about objects to be interpreted in the future. Therefore, a lot of effort has been done to formalize metadata standards such as Dublin Core[33], Pica-XML[34] and marc21[35]. But due to the fact that we interpret metadata as information units on its own it can also change over time. It is even very likely considering semantic techniques: Such techniques may evolve over time, increasing their quality, e.g. being able to identify more and more entities within a document. Thus, when talking about metadata quality, it is important not only to identify information units but also metadata records. The digital object identifier (DOI) system [110] takes account of this problem by being able to identify content objects in the digital environment. DOI names are assigned to any entity that is also metadata for use on digital networks. Summarizing the findings about the information life cycle we can state that

- we have to trace the quality of an information unit (including metadata records) over the whole life cycle
- we have to provide provenance information about the creation process of metadata records
- we must be able to provide fixity for information units to gain trust in the generated metadata

Finally, incorporating these findings into the information life cycle, we can introduce a high level model for the quality flow within semantic digital libraries (Fig. 36). During the creation and distribution stage, we will accumulate the outcomes of all quality metrics. In general, the quality metrics for each step are independent, but due to the fact, that 100% quality can only be achieved, without any errors in the used data, we have to consider all errors from the previous steps. Considering a normalized quality scale from 0 (0%) to 1 (100 %), we decided to multiply each quality value. Within the seeking stage, this product value can be used to visualize the quality to the user in addition to the individual quality of the semantic techniques as discussed in section 4.3.

[33] http://dublincore.org

[34] http://www.gbv.de/wikis/cls/PICA_XML_Version_1.0

[35] http://www.loc.gov/marc/bibliographic/ecbdhome.html

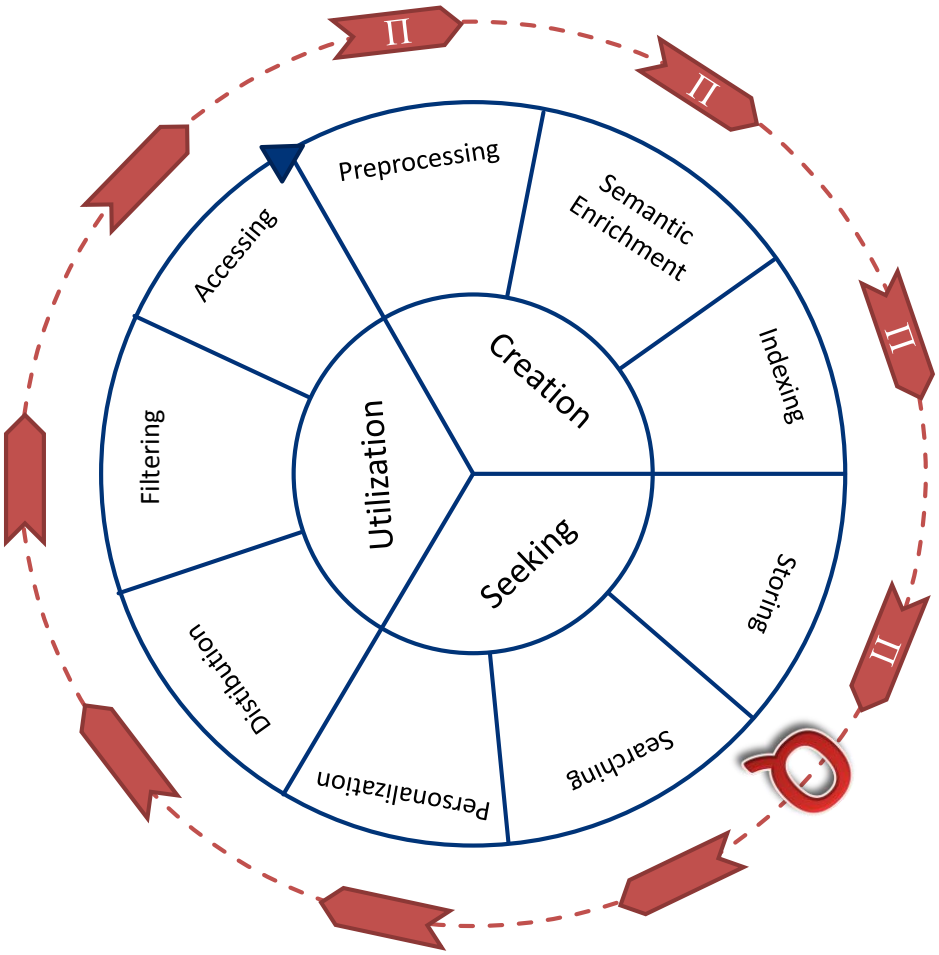


Fig. 36. Adapted Information Life Cycle with Quality Flow

Chapter 5

Lessons Learned: Deriving a Quality Model and its Application in Digital Libraries

In the previous chapter we have shown, that knowledge about the generation process of semantic enriched metadata is essential for the quality of a digital library. Furthermore, the usage of semantic technologies for automated metadata generation puts demands to the traceability and usage of such metadata. For example, we found out, that domain experts can judge the outcome of semantic technologies better, if the underlying quality is also visualized (see section 4.3). These results show that the old-established quality models are no longer sufficient.

Furthermore, with the growth of the Internet as a publication platform and especially with the growing number of open access journals and 'grey literature'/preprint servers, more and more high quality content is made available all the time. Beside the content that is provided by publishers known in a specific domain also this Web content is increasingly important for digital library users. Hence many digital libraries have extended their collections by harvesting content from the Web. However, unlike the controlled content that arrives from traditional publishers, assessing the quality of harvested content poses severe challenges. Whereas the general quality of each item or Web information source can usually be assessed quite well by the community of users (e.g. using feedback in Web 2.0 interfaces), the mission-critical problem of correctly indexing the new content for retrieval with both bibliographic data and content-based index terms remains with the digital library provider. Whereas gathering information from the Web has often been compared to 'trying to drink from a fire hydrant', indexing all this information with controlled quality seems like 'trying to drink from a fire hydrant while assessing the water quality of each sip'. This leads to a trade-off for digital libraries between offering broad and up-to-date document collections and providing high quality metadata for retrieval as we have shown in the last chapters.

A first step to collect at least the existing mostly bibliographic metadata from harvested sources is provided by the Protocol for Metadata Harvesting (OAI-PMH) of the Open Archive Initiative. OAI-PMH is a low-barrier mechanism for repository interoperability where content and service providers can share metadata. In particular it allows digital library providers to query the entire set of standardized metadata fields offered by each source. But as there are a lot of different content providers using different metadata formats, the quality of the respective values (particularly how they have been generated) is hard to assess. Still, the received metadata has to be transformed into a (global) metadata format used within

the library and is subsequently used indiscriminately. Moreover, even if the full text of some document can be exploited by a digital library provider for individually generating metadata (e.g., using heuristic or statistical methods), the provider currently has no possibility to illustrate the quality assessment to the end user.

In this chapter we will combine our lessons learned from the last chapters and make recommendations for action. Therefore, we will introduce the concept of lineage information for a time based component of the introduced quality model suitable for semantic digital libraries. The main difference between our proposed model and already existing quality models for digital libraries are twofold:

1. Our metadata quality model is utterly independent of any quality metric. For this reason it is applicable and adaptable for any information provider.
2. In conjunction with the time based approach we are able to track the quality of the metadata repository over time.

Afterwards, we will introduce a novel architecture for quality-controlled digital libraries similar to the staging area in data warehouses where the Extract Transform Load (ETL) process [111] takes place.

5.1. Lineage Information

In the last section, we have shown that information about the generation process of not only semantic metadata is essential and that in practice this kind of information is not available. To solve this problem, we recommend tracing the evolution of metadata records during their life cycle. Furthermore, this information should be globally available and machine readable to use it during automatic metadata enrichment and quality checks. This entails that we have to define a unique standard with the power to describe this kind of information.

It seems obvious to reuse an established concept from the domain of data warehousing the so called data lineage [112] or data provenance [113], [114]. Information stored in data warehouses is collected from many different sources and integrated as materialize views in local databases. Sometimes it is necessary to retrace the way from the original data sources to the aggregated data shown in the view. An administrator is able to learn why a specific tuple is shown in the respective view by tracing its lineage information to see all source tuples that have produced this view tuple. One can clearly see the similarities to our scenario. Also digital libraries collect data from different sources and integrate them into their repository. With the usage of lineage information, it would be possible to trace the evolution of metadata records. Fig. 37. visualizes our example workflow with integrated lineage information. In this way, we cannot only trace the evolution of the metadata on the horizontal axis (transformation steps), but also on the vertical axis (time). But how could such lineage information look like? All the more if we consider the different requirements of the different domains, e.g. chemistry, com-

puter science, architecture or mechanical engineering, do we have to confess that finding the "one size fits all" solution will not work?

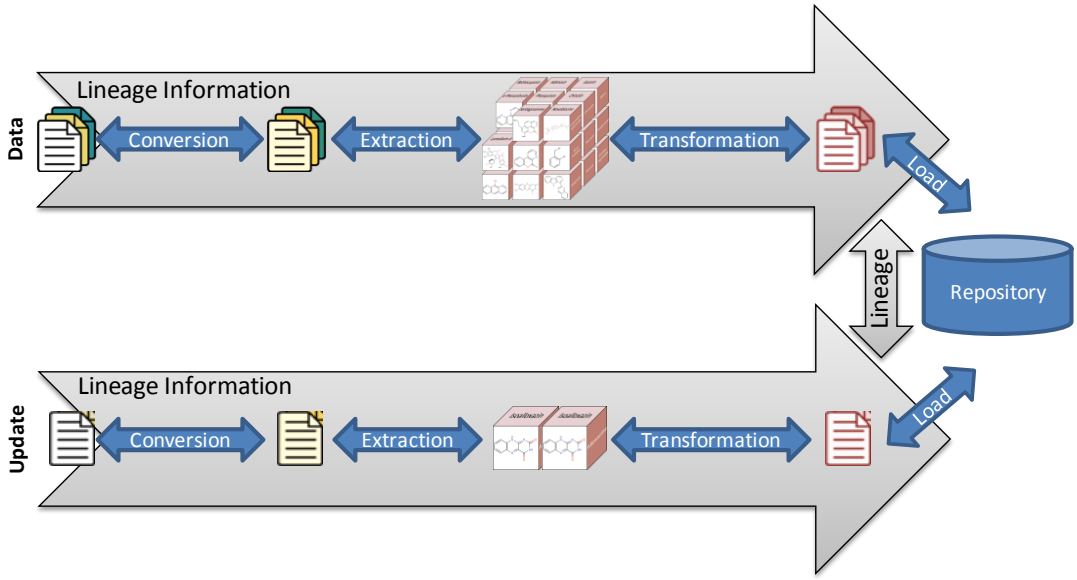


Fig. 37. Example Workflow with Lineage Information

Considering related work already done in the field of data provenance, we can see that this kind of information is always highly domain dependent. Thus, we can find literature about lineage models in specific domain, e.g. [115–117], or based on well-defined workflow engines, e.g. [118–120].

Emphasizing the main objective, i.e. the automatic quality assessment of metadata records, we may reduce this overwhelming problem of defining a global and domain spanning metadata lineage schema into a smaller one. We need a generic quality model which makes it possible to determine the quality of a specific metadata record in a specific version at any time.

5.2. Quality Model

As the main focus of this thesis is the quality control of a digital library and especially the quality of automatically enriched metadata, the focus of this section is the *describing and organizing* step of our life cycle. We have shown that data quality can be measured by several metrics and these metrics can be objective or subjective and are highly tasked dependent (see section 2.2). This implies that a quality model has to integrate the concept of a metric as well as the concept of individual requirements. Therefore, we will recommend a high-level metadata quality model, which is based on a generic data quality model from [121]. In the end, we will come up with a short description of how this model could also be adapted for the first and the last step within the *creation* phase of the lifecycle. Of course, the quality feedback in the *seeking* phase will also be considered in the model.

Our model (see Fig. 38) is divided into three areas: *local*, *local*/global, and *global*. The latter means, that the related model parts should be globally available because this part is universally valid as we will presently see. The local part includes information that is highly content dependent and should therefore be stored within the respective organization, i.e. digital library. For the model part assigned to *global*/*local* such a strict allocation is not possible. Here, we have to decide case-by-case. This separation will emerge in our example architecture in section 5.3.

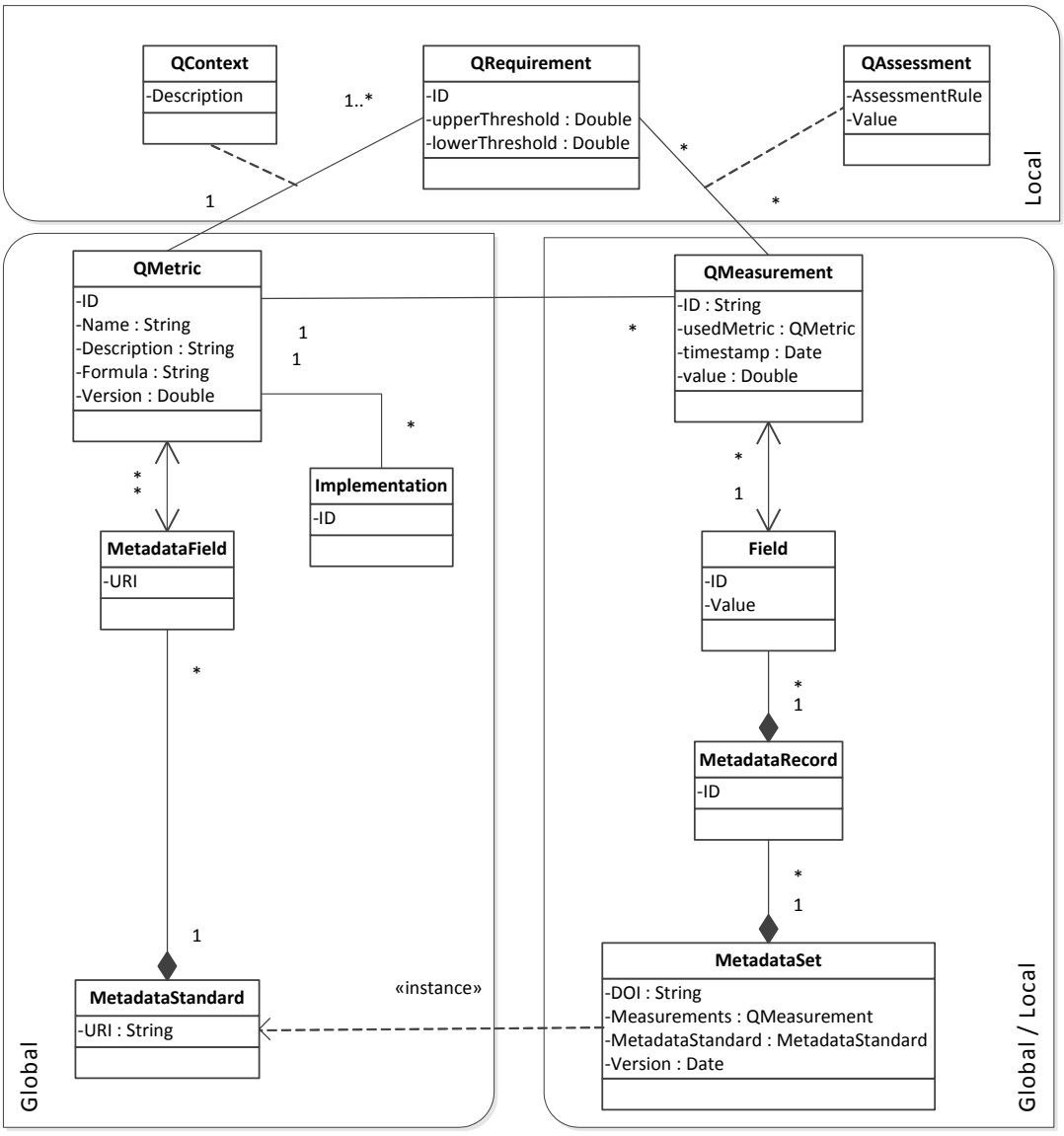


Fig. 38. Metadata Quality Model

Global When talking about digital libraries and document collections, we have to consider different *metadata standards*, e.g. Dublin Core (DC)[36], Pica-XML[37], MARC21[38] or Publishing Requirements for Industry Standard Metadata (PRISM)[39]. Taking also semantic metadata into account, the list of possible standards gets even longer, e.g. SciXML [122] in the chemical domain. Every single one of them is uniquely identifiable by its URI, and defines several metadata fields, e.g. "title", "creator", or "abstract". Each of these fields has a unique identifier, i.e. its URI.

As already mentioned, the contents vary and so does the quality assessment from field to field. For instance, in Dublin Core, the "identifier" field may use a metric analyzing the existence of such an identifier, whereas the "date" metric may check the validity of the given date. Therefore, it must be possible to assign different metrics to each individual *metadata field* and not one "global" metric for a whole metadata standard. In our model this is achieved by a quality metric instance associated with a metadata field. Such a quality metric (*QMetric*) is defined as a quantifiable inspection criterion. Therefore, each QMetric has to have a name, a verbal description, the actual formula, which should be stored in a TEX format, and versioning information. Using TEX, it is quite trivial to interpret and calculate the formula in a later step. Due to compatibility reasons, the outcome of each metric should be normalized to [0,1], meaning a quality of 0% up to 100%. In addition we can also store references to concrete implementations of the metric.

Global/Local In a next step, we need to measure the quality for individual instances of the metadata standard, i.e. metadata set. This set can differ depending on the version. Thus, it is essential for the model, that each set in a specific version can be identified. This can be achieved by assigning a digital object identifier (DOI®) to each set.

For example, the TIB includes data from the Beilstein Journal of Organic Chemistry (BJOC)[40] and the Archive of Organic Chemistry (ARKIVOC)[41], which is one of the oldest open access journals in Organic Chemistry. While ARKIVOC does not provide any metadata information based on a metadata standard, BJOC provides well-structured metadata information based on DC, PRISM and an own metadata standard. We will focus on the DC metadata. Thus, we crawl all records available from BJOC at a specific time (which is our version) and extract the DC related metadata. For this set, we assign a DOI and create an entry in our model. Afterwards, we create an entry for each record and its related fields, assigning the URI

[36] http://dublincore.org/documents/2010/10/11/dcmi-terms

[37] http://www.gbv.de/wikis/cls/PICA_XML_Version_1.0

[38] http://www.loc.gov/marc/bibliographic

[39] http://www.prismstandard.org/specifications

[40] http://www.beilstein-journals.org/bjoc

[41] http://www.arkat-usa.org

as identifier, the value, and a connection to its record and thus to its underlying metadata standard.

Now, it is possible to perform a measurement for a specific metric in relation to a metadata set. The reference to the used metric is stored within the *QMeasurement* object. Each QMeasurement has been performed at a specific timestamp and the outcome is a value between [0,1] as we defined in QMetric. Based on these concepts, we will be able to specify the overall quality of a metadata set S. The metadata set is defined as a set of metadata records r_i:

$$S = \{r_1, \dots, r_i, \dots, r_n\} \tag{1}$$

Each metadata record r_i is defined as the set of fields f_j describing the metadata:

$$r_i = \{f_1, \dots, f_j, \dots, f_m\} \tag{2}$$

Given a quality measurement Q for a specific field f_j and an importance factor α for that field, the overall quality of a metadata record can be expressed by equation (3). Such an importance factor is essential, because a library may stress the quality of a specific field, whereas other fields of the record are less important for their retrieval.

$$Q_{r_i} = \frac{\sum_{j=1}^{m} \alpha * Q_{f_j}}{|r_i|} \tag{3}$$

Resulting in an overall quality of the whole metadata set:

$$Q_S = \frac{\sum_{i=1}^{n} Q_{r_i}}{|S|} \tag{4}$$

As we could see a metadata set will evolve over time. Updates will be provided and each update will have its own quality. Thus, we need *quality lineage information* to judge the overall quality of a metadata set over time. Given equation (4) as the quality of a metadata set to a given time t_0 we can determine the quality after an update at t_n with:

$$\frac{\sum_{i=0}^{n} Q_{S_{t_i}} * \left|S_{t_i}\right|}{\sum_{i=0}^{n} \left|S_{t_i}\right|} \tag{5}$$

Local After we can globally quantify the quality of a specific metadata set, we have to judge this value in the context of a specific digital library. The aim is to be able to visualize the quality to a librarian or to a user. Here, we will use the well know *traffic light paradigm*: a good quality will be visualized green, a middle quality yellow and a bad quality red. Therefore, we need the following parts of our model. By means of *QRequirement* the quality level that a QMetric should have for a metadata field is specified. Typically it defines a set of thresholds (upper and lower) which will be used during the quality assessment. It is important that each QRequirement is only valid for a specific context which is verbally described. In the domain of digital libraries such contexts could be, for example, the kind of end user such as student, professor or business user. In this way we are able to accommodate dif-

ferent users or decision making contexts that have different criteria for assessing the data quality levels that are appropriate for their particular task.

After we defined quality requirements for a specific context and we conducted some measurements, we are ready to apply our assessment rules. A *QAssessment* is the result of applying such an assessment rule. Normally it is the value that compares a measurement to a requirement. Therefore, the outcome (value) of the measurement is somehow judged regarding the requirement thresholds and will result in a "traffic light value".

Utilizing this model, we are now able to specify and reference one or more metrics for a specific metadata field. In addition, each information provider can use this knowledge to actually perform measurements of owning metadata sets and judge the quality according to their requirements. In the next section, we will incorporate this model into a novel architecture suitable to use in a digital library workflow.

5.3. Example Architecture

For finding a suitable architecture where we can incorporate our quality model we analyzed the workflow of the TIB, described in the introducing motivation.

1. The TIB receives metadata sets from several information providers based on different metadata standards.
2. For each standard they define a transformation process, to convert the provided standard into their own standard. Each transformation step is audited once during its creation.
3. Each provided set is transformed record by record.
4. The imported and updated records are evaluated according to their completeness.

This process is similar to the well-known architecture pattern from the area of data warehousing, named Extract, Transform and Load (ETL). ETL is an industry standard term used to represent the data movement and transformation processes. As the name suggests this concepts consists of three different parts.

Extraction: This part deals with the extraction of data from one or more relevant heterogeneous data sources. An intrinsic part of the extraction involves the parsing of the extracted data, resulting in a check if the data meets a needed structure. If this quality check fails, the data may be already rejected entirely or in part. In general, the goal of the extraction phase is to convert the metadata into a single format which is appropriate for the transformation process.

Transformation: The transformation applies a series of business rules or functions to the extracted data to derive the data to fit the operational needs. Some data sources will require very little manipulating and in other cases, some transfor-

mations like cleansing, standardization, aggregation or quality checks may be required.

Loading: The load phase loads the data into the operational database. Depending on the requirements, this process can vary a lot: some providers may overwrite the old data without any logging, whereas others can maintain a history and audit trail of all changes to the data loaded into the operational database. However, as the load phase directly interacts with the operational database, also the constraints defined in the database schema, e.g. trigger, make a contribution to the overall data quality performance of the ETL process.

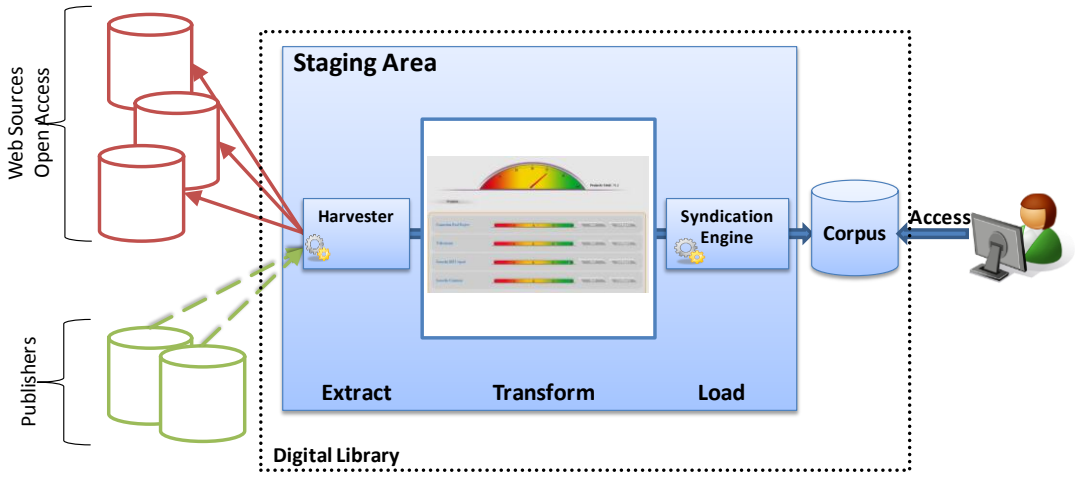


Fig. 39. High-level View on the ETL Architecture

Adapting these general steps to the digital library domain (cf. Fig. 39) the first step involves the consolidation of the metadata from different content providers. Each separate provider may use its own metadata format. The digital library provider collects all metadata from the different providers and stores them in a temporary data store (extract). The metadata can either be transmitted using a push mechanism meaning that the metadata providers are responsible for delivering their data to the digital library. The other possibility is to use a pull process meaning that the digital library requests the data from the providers. Therefore, in our architecture the data transmission is based on the OAI-PMH protocol. The advantage of this protocol is the required mapping of the metadata to a small subset of the Dublin Core format and the possibility to transfer additional, well-described metadata formats. Thus, it is easy for the digital library to parse, validate and extract the needed metadata. During the transformation phase, the digital library will convert the variety of external schemas into their own schema and load the resulting metadata into their operational corpus, if the resulting quality meets their needs. This means, that the outcome of a QAssessment has a value defined in a QRequirement.

In this way, we have a solid framework for a well-defined workflow in a digital library and regarding the implementation of this framework every library can already rely on best practices from the domain of data warehousing. But as already mentioned before, we still have to assess the quality of the metadata.

5.3.1. Quality Control

In the previous sections we have shown, that it is essential to have background knowledge about generated metadata sets to determine their quality. Thus, the most important part of our extended ETL architecture (see Fig. 40) is the staging area with access to a repository with quality information about the metadata set. The information stored in this repository is compliant with our model described in section 5.2.

The universal information about metadata standards, their fields, and the respective quality metrics should be stored in a centralized repository. At least, the access to such information should be centralized. Therefore, it is easily conceivable to launch a service similar to the DOI resolver system (we call it *QCite*) where such information is stored, respectively linked.

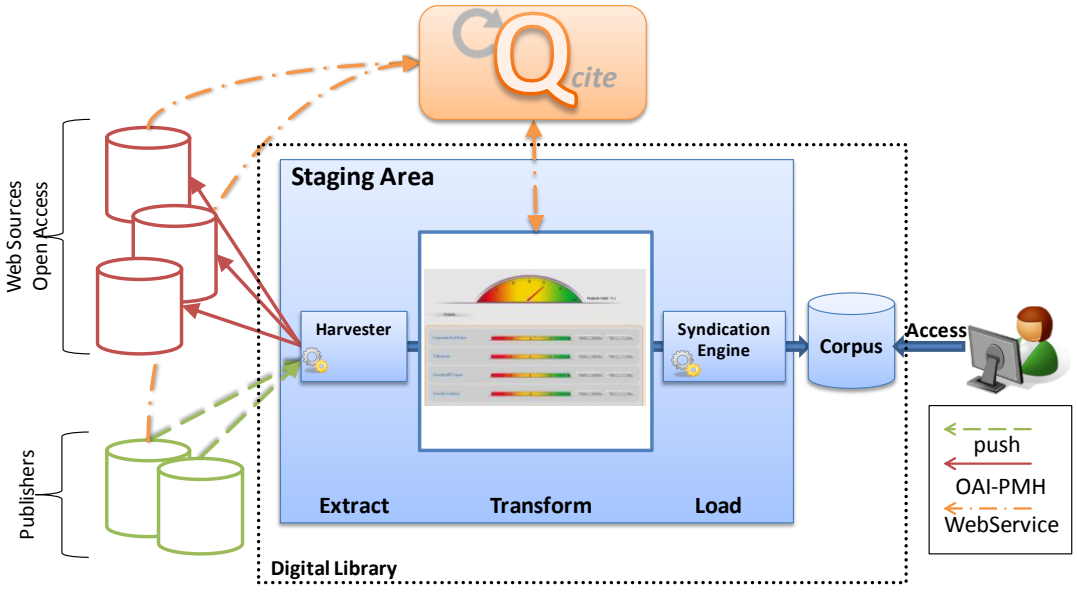


Fig. 40. High-level View on the ETL Architecture with Access to Quality Information

In contrast, the quality information about the metadata sets should always be delivered by the information providers together with the metadata. This could be done by providing a RDF format to express our model information or by providing a well-defined interface to some Web service running at the provider's side (we choose the later, as we will see later). In this way, the library can rely on the given

information or decide to recalculate the quality information based on some metrics provided by the QCite repository. This information should again be hold available for other consumers. In this scenario, everybody could determine the quality of a specific metadata set without having access to the (maybe restricted) content. Still, a library can decide if the underlying metadata quality meets its requirement and could store this information in its local repositories. This part should be integrated in the in-house ETL architecture.

Coming back to our use case, Fig. 41 shows a usual workflow for interacting with our system. First the TIB retrieves a metadata set of chemical documents from a publisher. As additional information, the TIB retrieves the DOI of this set, which can be used to enquire the QCite repository for quality information about this metadata set. They retrieve a list of measurements executed on this set, selecting the measurement using the "completeness metric". As a requirement, the TIB defined, only using metadata sets, where all DC metadata fields are filled, which is valid for this set. Thus, the system confirms the assessment and the workflow can go on. As this metadata also contains semantic metadata, i.e. chemical entities, the TIB has the requirement, that they only provide chemical entities also available in the PubChem database. Unfortunately, this quality check is not done by the provider and thus has to be performed within the TIB. After performing this measurement, the TIB will submit the result back to the QCite repository. In the meantime, the system already assesses the TIB requirement, whether the recognition rate is between 90% and 100%. After retrieving a positive outcome, the automatic transformation into the internal schema starts. Finally, the TIB checks the transformed metadata set against the completeness metric, stores the quality information into its internal repository and loads the data into the operational database, as the quality assessment was positive.

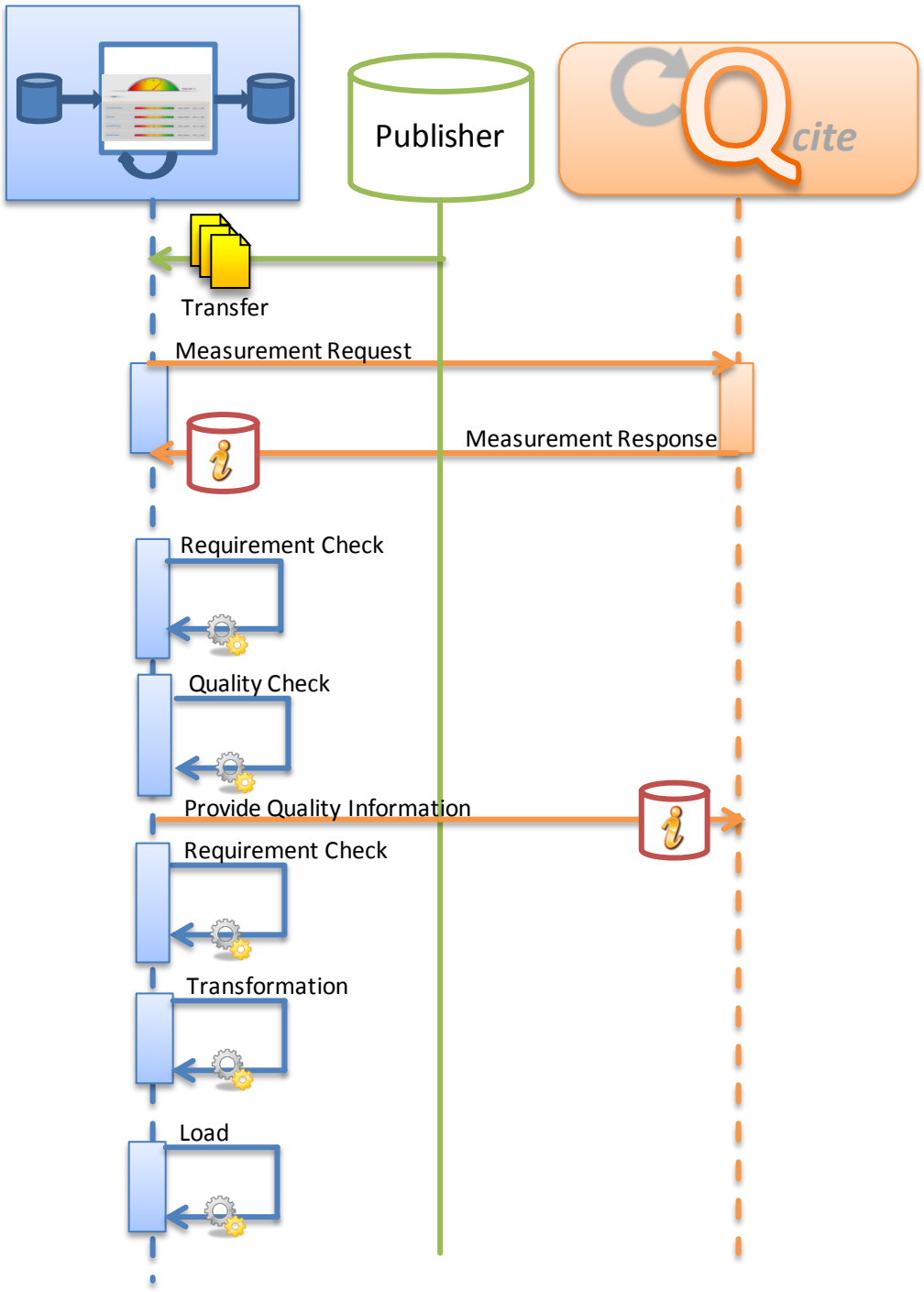


Fig. 41. A Sequence Diagram describing the Meta-Line Workflow.

5.3.2. Why Web Services? An Experiment

The decision for using Web services as communication elements in our architecture is twofold:

1. We described in the last subsection that the needed information is spread around the global players, i.e. "*QCite*", libraries and content providers. As a consequence we need a well-defined interface.
2. As we will see later in a small experiment, the amount of information handled by digital libraries requires a distributed architecture.

We conducted an experiment and logged the amount of metadata, which has been processed in the context of GetInfo[42]. Fig. 42 shows the total amount of metadata processed by the TIB within the last 18 months. In total, 48.991.083 entries have been processed, resulting in an average of 89.481 entries per day.

In addition, we implemented a prototype of our architecture. For the implementation we used Java as programming language and Axis2 as Web service framework. The quality information is stored in a MySQL database. We filled our QCite repository with example data containing quality information about 120 Beilstein Journal metadata records, each of them containing 6 DC fields, i.e. title, abstract, identifier, rights, format, and language. Finally, we deployed the QCite service on a Tomcat server. The interface description can be found in the WSDL file[43].

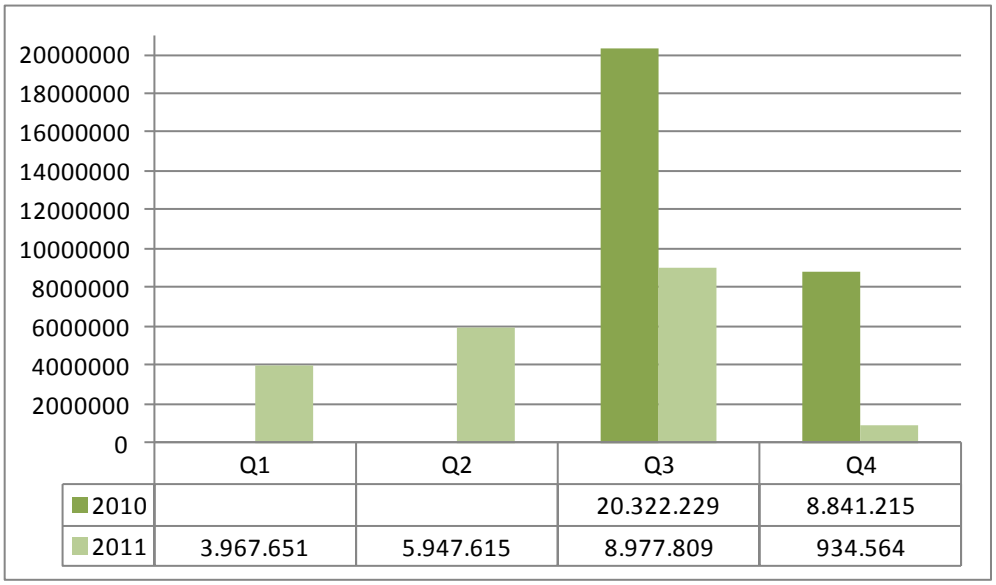


Fig. 42. Number of metadata entries totally processed per quarterly period

[42] https://getinfo.de

[43] http://tbdb.l3s.uni-hannover.de:8080/QCiteServer/services/QCiteWs?wsdl

We performed 1000 measurements, each measurement consist of 720 service calls, representing quality checks for 120 metadata records consisting of 6 metadata fields. The measurements have been performed within the local network and outside the local network, so that we can predict a mean value. As we can see in Fig. 43 the time varied from 468 milliseconds to 15 milliseconds, resulting in an overall average of about 56 milliseconds. If we take this average and multiple it with the number of entities processed by the TIB per day (89.481), assuming 6 fields each, we would need around 8 hours per day for the quality checkups. This is a reasonable time, if we consider that the daily update process has to be done within 24 hours. Of course, the usage of aggregation and analytic functions would further reduce this time. One example would be the aggregation of quality values for one field per metadata record or metadata set. Using an aggregated quality value for a metadata record, we would reduce the number of queries to 120 and the overall process time would be reduced to 83 minutes. Using the aggregation for the whole metadata set we would further reduce this amount to one single query and a process time of less than 1 second. Of course, we would have to add processing time for the aggregation itself but this is negligible. Thus we can make a tradeoff between quality granularity and performance.

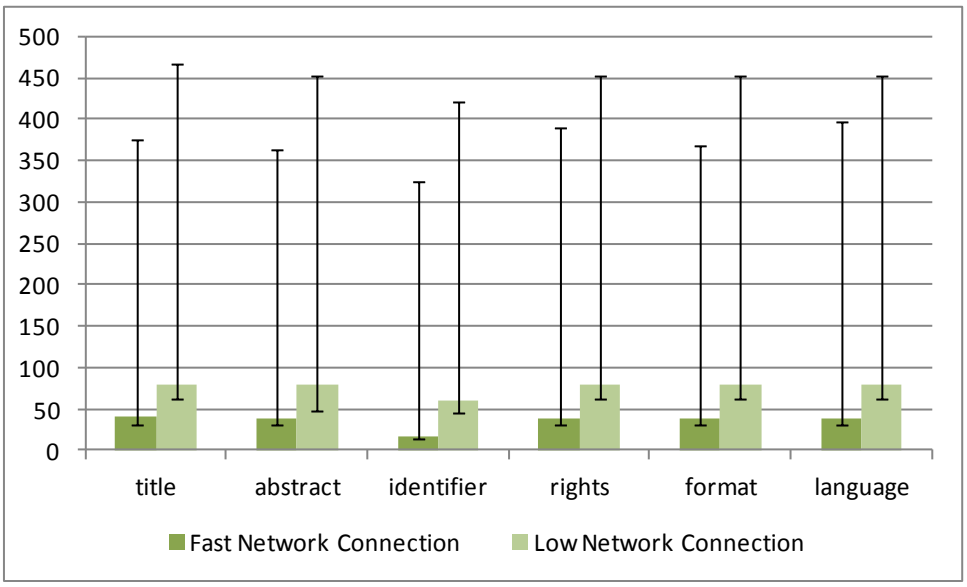


Fig. 43. Average time in milliseconds for a Web service call with error rates

5.4. Conclusion

In this chapter we discussed the concept of data lineage, i.e. the ability of visualizing data in a reverse order, following its transformation step by step. We illustrated, that it seems unlikely to come up with a domain spanning lineage model, but our main objective, the automatic quality assessment can still be solved by introducing our fine-grained quality model.

Our recommended quality model distinguishes between the general metadata standard and concrete instances of this standard. Therefore, we can define several quality metrics for each field of the standard and we can perform concrete measurements of metadata records based on these metrics. In addition, each library can define its own requirements and assess them. Based on this model, we defined an equation to determine the overall quality of a metadata record over time.

Finally, we adopted the well-known ETL workflow from the area of data warehousing and combined it with a central quality repository. By using a centralized repository, we are able to reuse measurements, thus the quality assessment is faster. This repository can be integrated into a workflow, to evaluate the metadata sets and automatically determine their quality. To evaluate this architecture, we performed an experiment showing that the time needed for 536.886 service calls is reasonable low (around 8 hours). Of course this time can be further reduced by the usage of aggregation functions and by parallelization.

Chapter 6

Conclusion and Future Work

6.1. Summary of Contributions

One of the main problems of digital libraries today is to find a suitable trade-off between up-to-date document collections covering also a growing amount of online content and sufficient data quality achieving the high quality standards of a library. Beside bibliographic metadata also semantic, domain-specific metadata is important to assist consumers with their information gathering process offering specialized search interfaces or drill-down facets. However, due to the well-known problem of information overload it is not possible to manually assess the quality of all generated metadata. Therefore, it is mandatory to make the automatic generation workflows transparent in order to satisfy the high quality standards of libraries. But, especially when using semantic techniques this is a challenging task because sometimes even the outcome of the technique is not easily assessable. Motivated by the work within the ViFaChem project in tight cooperation with the German National Library of Science and Technology Hannover (TIB) we investigated this problem in this thesis in detail.

In chapter 2, we discussed the general quality concept of digital libraries. We found out that quality of digital libraries is still defined conservative in the sense of user satisfaction. Until now, only a very limited amount of research has been done to judge the quality of automatically generated metadata. For semantic metadata which is usually very domain specific, the amount of research is even less. Hence, we proposed the need of a higher-level quality model, in which different quality metrics can be chosen, based on the underlying technology and metadata standard. Furthermore, the model still allows a statement about the quality of the resulting metadata. For illustration purposes we introduced a use case scenario, i.e. chemical digital libraries, in the chapter 3.

Based on the information life cycle we further identifying where and when quality indicators can be measured: preprocessing, semantic metadata enrichment, indexing, and document retrieval.

We have further shown that the underlying creation process of the metadata has a high influence on the retrieved quality. Furthermore, we have shown that also the semantic meaning of a metadata field needs consolidation. Therefore, we conducted a user study in the field of chemistry observing some experts' interaction with automatic created metadata. The study resulted in three major observations which have than been transformed into three general quality metrics namely

degree of category coverage (DCC), semantic word bandwidth (SWB) and relevance of covered terms (RCT). Furthermore, we investigated the usefulness of the proposed metrics and their information gain. We showed that the domain experts were easily able to assess the outcome of the technique and gained insights into what quality to expect during their information gathering. Consequently it is indeed useful to measure the quality of a semantic technique.

Furthermore, digital libraries have to face the *problem of the hidden Web*. That means, most of the indexes build by digital libraries are not detectable by standard Web search engines and thus are hidden to most users because a simple Web search is the standard way of information seeking in the Web today. We have shown one solution in the domain of chemistry to open up these indexes. We developed a workflow allowing for automatic generation of customized index pages including all metadata information extracted from publicly accessible databases for each occurring chemical entity. Our framework can easily be used, e.g., by libraries, open access journals, or other content providers in the chemical domain. We also performed experiments to show the usefulness of our approach. The retrieval quality of our enriched index pages is almost as good as chemical exact structure searches and significantly better compared to a baseline/full text search.

Moreover, we have shown that assessing the quality of semantic techniques is tasked dependent. Indeed, the way a task is defined is domain specific and leads to the general problem of assessing the similarity of entities. As a result, digital libraries have to take this task dependence into account and have to assess the usefulness of similarity in relation to the audience and their task. In the discourse of this thesis, we analyzed several similarity measures in the domain of chemistry and based on the findings, we developed a personalized information system for the chemical domain.

Finally in chapter 5 we combined our findings to make recommends for action. We discussed the concept of data lineage process, i.e. the ability of visualizing data in a reverse order, following its transformation step by step. We could show, that it seems unlikely to come up with a domain spanning lineage model, but our main objective, the automatic quality assessment can still be solved by introducing our fine-grained quality model. Our introduced quality model distinguishes between the general metadata standard and concrete instances of this standard. Therefore, we can define several quality metrics for each field of the standard and we can perform concrete measurements of metadata records based on these metrics. In addition, each library can define its own requirements and assess them. Based on this model, we defined an equation to determine the overall quality of a metadata record over time.

6.2. Open Directions

Let us conclude with a short outlook. Quality in digital libraries is an integral part of their services. Without this underlying high quality metadata their economic survival will be questionable. For this reason, they are still manually assessing documents and generating knowledge in the sense of metadata. However, due to the information flood this is not really feasible resulting in the separation of topical centered digital libraries, e.g. German National Library of Science and Technology (TIB), German National Library of Medicine (ZBMed), and Leibniz Information Centre for Economics (ZBW). However, also for each of these topic centered digital libraries the amount of information cannot be handled anymore in a manual way. Consequently even in the future it is important to put more effort in the development of quality metrics for semantic techniques. A good starting point will be our procedure observing domain experts while operating with the outcome of the respective semantic technique.

Still, based on the success in the field of the community based creation of ontologies [123], [124] and metadata standards [125], the digital library community should put effort into the establishment of a community based portal for quality metrics. The functionality could be similar to the functionality of the myExperiment portal[44]. That implies building domain dependent communities, finding, sharing, and creating standards for semantic metadata and its related metrics.

Another important aspect is the formalization of the underlying protocol to be able to easily integrate this quality portal into the workflow of a digital library. Of course, it would be a good way to integrate this idea into the already standardized and accepted OAI-PMH protocol.

[44] http://www.myexperiment.org/

List of Figures

Bibliography

[1] Thomson Reuters, "Web of Science," 2011. [Online]. Available: http://wokinfo.com/products_tools/multidisciplinary/webofscience/. [Accessed: 25-Jul-2011].

[2] Elsevier Inc., "Engineering Village," 2011. [Online]. Available: http://www.engineeringvillage.com. [Accessed: 25-Jul-2011].

[3] Elsevier Inc., "SciVerse Scopus," 2011. [Online]. Available: http://www.info.sciverse.com. [Accessed: 25-Jul-2011].

[4] Technische Informationsbibliothek, "GetInfo." [Online]. Available: http://www.getinfo.de. [Accessed: 25-Jul-2011].

[5] Cas-pr@cas.org, "CAS REGISTRY, the global standard for chemical research, approaches 50 millionth registration milestone," 2009. [Online]. Available: http://www.cas.org/newsevents/releases/casregistry50m081609.html. [Accessed: 25-Jul-2011].

[6] Deutsche Forschungsgemeinde, "Wissenschaftliche literaturversorgungs- und Informationssysteme: Schwerpunkt der Förderung bis 2015." Bonn, Germany, p. 9, 2006.

[7] Bund-Länder Kommission, "Neuausrichtung der öffentlich geförderten Informationseinrichtungen," Bonn, 2006.

[8] T. Risse, P. Knezevic, C. Meghini, R. Hecht, and F. Basile, "The BRICKS infrastructure - An Overview," in *Proc. of 75th Conference on Electronic Imaging, the Visual Arts & Beyond (EVA)*, 2005.

[9] S. R. Kruk, T. Woroniecki, and A. Gzella, "JeromeDL – a Semantic Digital Library," in *Proceedings of the Semantic Web Challenge*, 2007.

[10] L. Candela, D. Castelli, P. Pagano, C. Thano, Y. Ioannidis, G. Koutrika, S. Ross, H.-J. Schek, and H. Schuldt, "Setting the foundations of digital libraries," *D-Lib Magazine*, vol. 13, no. 3/4, 2007.

[11] J. Beall, "Metadata and data quality problems in the digital library," *Journal of Digital Information*, vol. 6, no. 3, p. 355, 2005.

[12] S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Building Chemical Information Systems - the ViFaChem II Project," in *Proc. of Datenbanksysteme in Business, Technologie und Web (BTW)*, 2009.

[13] X. Ochoa and E. Duval, "Automatic evaluation of metadata quality in digital repositories," *International Journal on Digital Libraries*, vol. 10, no. 2–3, pp. 67–91, Aug. 2009.

[14] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp, "Streams, structures, spaces, scenarios, societies (5s)," *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 270–312, Apr. 2004.

[15] T. Saracevic, "Digital library evaluation: toward evolution concepts," *Library Trends*, vol. 49, no. 2, pp. 350–369, 2000.

[16] P. Constantopoulos, I. Sølvberg, N. Fuhr, P. Hansen, M. Mabe, and A. Micsik, "Digital Libraries: A Generic Classification and Evaluation Scheme," in *Research and Advanced Technology for Digital Libraries, 5th European Conference, ECDL 2001*, 2001, vol. 2163, pp. 187–199–199.

[17] M. A. Gonçalvesa, B. L. Moreira, E. A. Fox, and L. T. Watson, "'What is a good digital library?' – A quality model for digital libraries," *Information Processing & Management*, vol. 43, no. 5, pp. 1416–1437, Sep. 2007.

[18] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Sølvberg, "Evaluation of digital libraries," *International Journal on Digital Libraries*, vol. 8, no. 1, pp. 21–38, Feb. 2007.

[19] M. Khoo, J. Pagano, A. L. Washington, M. Recker, B. Palmer, and R. A. Donahue, "Using web metrics to analyze digital libraries," in *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries - JCDL '08*, 2008, pp. 375–384.

[20] S. R. Kruk, E. Kruk, and K. Stankiewicz, "Evaluation of Semantic and Social Technologies for Digital Libraries," in *Research and Advanced Technology for Digital Libraries*, 2008, pp. 74–77.

[21] J. Beall, "Metadata and data quality problems in the digital library," *Journal of Digital Information*, vol. 6, no. 3, p. 355, 2005.

[22] J. Barton, S. Currier, and J. M. N. Hey, "Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice," in *International conference on Dublin Core and metadata applications: supporting communities of discourse and practice--- metadata research & applications*, Seattle, Washington: , 2003, pp. 39–48.

[23] L. Al-Hakim, *Information quality management: theory and applications*. Idea Group Publishing, 2007.

[24] X. Liu, K. Maly, M. Zubair, and M. Nelson, "Arc-An OAI Service Provider for Digital Library Federation," *D-Lib Magazine*, vol. 7, no. 4, 2001.

[25] S. E. Thomas, "Quality in bibliographic control," *Journal of Library Trends*, vol. 44, no. 3, pp. 491–505, 1996.

[26] J. Greenberg, M. C. Pattuelli, B. Parsia, and W. D. Robertson, "Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization," *Journal of Digital Information*, vol. 2, pp. 38–46, 2001.

[27] W. E. Moen, E. L. Stewart, and C. R. McClure, "Assessing Metadata Quality: Findings and Methodological Considerations from an Evaluation of the U.S. Government Information Locator Service (GILS)," p. 246, Apr. 1998.

[28] S. L. Shreeves, E. M. Knutson, B. Stvilia, C. L. Palmer, M. B. Twidale, and T. W. Cole, "Is 'quality' metadata 'shareable' metadata? The implications of local metadata practices for federated collections."In: Proceedings of the Association of College and Research Libraries (ACRL) 12th National Conference. Minneapolis, MN, pp. 223–237, 2005.

[29] B. Stvilia, L. Gasser, and M. B. Twidale, "Metadata Quality Problems in Federated Collections," in *Challenges of Managing Information Quality in Service Organizations*, L. Al-Hakim, Ed. Idea Group Publishing, 2007, pp. 154–186.

[30] A. J. Wilson, "Toward Releasing the Metadata Bottleneck," *ALCTS Newsletter*, vol. 51, no. 1, pp. 16–28, 2007.

[31] Y. Bui and J.-R. Park, "An assessment of metadata quality: A case study of the National Science Digital Library Metadata Repository," *Proceedings of CAISACSI 2006 Information Science Revisited Approaches to Innovation*, p. 13, 2006.

[32] B. Hughes, "Metadata Quality Evaluation: Experience from the Open Language Archives Community," in *Digital Libraries: International Collaboration and Cross-Fertilization*, vol. 3334, Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, and E. Lim, Eds. Springer Berlin / Heidelberg, 2005, pp. 135–148.

[33] J. Najjar, S. Ternier, and E. Duval, "User Behavior in Learning Objects Repositories: An Empirical Analysis." Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 4373–4378, 2004.

[34] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," *Communications of the ACM*, vol. 40, no. 5, pp. 103–110, 1997.

[35] X. Zhu and S. Gauch, "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web," *Proceedings*

of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 00, pp. 288–295, 2000.

[36] S. Ede, "Fitness for purpose: The Future Evolution of Bibliographic Records and Their Delivery," *Catalogue & Index*, no. 116, pp. 1–3, 1995.

[37] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith, "A framework for information quality assessment," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1720–1733, Oct. 2007.

[38] T. Bruce and D. Hillmann, "The continuum of metadata quality: defining, expressing, exploiting," in *Metadata in Practice*, D. I. Hillmann and E. L. Westbrooks, Eds. 2004, pp. 238 – 256.

[39] M. Sanderson and B. Croft, "Deriving concept hierarchies from text," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, 1999, pp. 206–213.

[40] P. Cimiano, S. Handschuh, and S. Staab, "Towards the self-annotating web," in *Proceedings of the 13th conference on World Wide Web - WWW '04*, 2004, p. 462.

[41] M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," in *Proceedings of Conference on Computational Linguistics (COLING)*, 1992, pp. 539–545.

[42] U.S. National Library of Medicine, "Fact Sheet MEDLINE." [Online]. Available: http://www.nlm.nih.gov/pubs/factsheets/medline.html. [Accessed: 13-Jul-2011].

[43] J. Diederich and W.-T. Balke, "Automatically created concept graphs using descriptive keywords in the medical domain.," *Methods of information in medicine*, vol. 47, no. 3, pp. 241–50, Jan. 2008.

[44] J. Diederich and W.-T. Balke, "The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems," in *Research and Advanced Technology for Digital Libraries (ECDL)*, 2007, vol. 4675, pp. 1–13.

[45] S. A. Golder and B. A. Huberman, "The Structure of Collaborative Tagging Systems," *CoRR - Computing Resource Repository*, 2005.

[46] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 211–220.

[47] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proceedings of the international conference on World Wide Web*, 2007, pp. 501–510.

[48] S. Chan, "Tagging and Searching – Serendipity and museum collection databases," in *Museums and the Web 2007*, 2007.

[49] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information Retrieval in Folksonomies: Search and Ranking," in *The Semantic Web: Research and Applications*, 2006, vol. 4011, pp. 411–426.

[50] K. Razikin, D. H.-L. Goh, A. Y. Chua, and C. S. Lee, "Can Social Tags Help You Find What You Want?," in *Research and Advanced Technology for Digital Libraries*, 2008, vol. 5173, pp. 50–61.

[51] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?," in *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 2008, pp. 193–202.

[52] R. Krestel and L. Chen, "The Art of Tagging: Measuring the Quality of Tags," in *3rd Asian Semantic Web Conference, ASWC*, 2008, vol. 5367, pp. 257–271.

[53] S. Tönnies and W.-T. Balke, "Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation," in *13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009*, 2009, pp. 168–179.

[54] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann, "Qood grid: A metaontology-based framework for ontology evaluation and selection," in *Proceedings of Evaluation of Ontologies for the Web, 4th International EON Workshop*, 2006.

[55] A. Lozano-Tello and A. Gómez-Pérez, "ONTOMETRIC : A Method to Choose the Appropriate Ontology," *Journal of Database Management*, vol. 15, no. 2, pp. 1–18, 2004.

[56] S. Tartir, B. Arpinar, M. Moore, A. Sheth, and B. Aleman-Meza, "OntoQA: Metric-Based Ontology Quality Analysis," in *Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*, 2005.

[57] D. Vrandečić and Y. Sure, "How to Design Better Ontology Metrics," in *The Semantic Web: Research and Applications*, 2007, pp. 311–325.

[58] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4ve, pp. 211–218, 2002.

[59] J.-R. Park, "Metadata Quality in Digital Repositories: A Survey of the Current State of the Art," *Cataloging Classification Quarterly*, vol. 47, no. 3, pp. 213–228, 2009.

[60] E. Leistner, "[Drugs from nature. The biology of taxane].," *Pharmazie in unserer Zeit*, vol. 34, no. 2, pp. 98–103, Jan. 2005.

[61] B. Sun, Q. Tan, P. Mitra, and C. L. Giles, "Extraction and search of chemical formulae in text documents on the web," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 251–260.

[62] B. Sun, P. Mitra, and C. L. Giles, "Mining, indexing, and searching for textual chemical molecule information on the web," in *Proceeding of the international conference on World Wide Web*, 2008, pp. 735–744.

[63] P. Murray-Rust and H. S. Rzepa, "Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles," *Journal of Chemical Information and Modeling*, vol. 39, no. 6, pp. 928–942, Nov. 1999.

[64] R. Hoffmann and P. Laszlo, "Representation in Chemistry," *Angewandte Chemie International Edition in English*, vol. 30, no. 1, pp. 1–16, 1991.

[65] J. R. McDaniel and J. R. Balmuth, "Kekule: OCR-optical chemical (structure) recognition," *Journal of Chemical Information and Modeling*, vol. 32, no. 4, pp. 373–378, Jul. 1992.

[66] A. T. Valko and a P. Johnson, "CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition.," *Journal of chemical information and modeling*, vol. 49, no. 4, pp. 780–7, Apr. 2009.

[67] M. Zimmermann, L. T. Bui Thi, and M. Hofmann, "Combating Illiteracy in Chemistry: Towards Computer-Based Chemical Structure Reconstruction," *ERCIM News*, vol. 60, pp. 40–41, 2005.

[68] I. V. Filippov and M. C. Nicklaus, "Optical structure recognition software to recover chemical information: OSRA, an open source solution.," *Journal of chemical information and modeling*, vol. 49, no. 3, pp. 740–743, Mar. 2009.

[69] P. Corbett and P. Murray-Rust, "High-Throughput Identification of Chemistry in Life Science Texts," in *Computational Life Sciences II*, 2006, vol. 4216, pp. 107–118.

[70] J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman, and P. Murray-Rust, "Chemical documents: machine understanding and automated information extraction.," *Organic & Biomolecular Chemistry*, vol. 2, no. 22, pp. 3294–300, Nov. 2004.

[71] H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.," *Journal of Chemical Documentation*, vol. 5, no. 2, pp. 107–113, 1965.

[72] D. J. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont.," *Journal of Chemical Documentation*, vol. 5, no. 1, pp. 43–51, Feb. 1965.

[73] E. G. Smith, *The Wiswesser Line-Formula Chemical Notation (WLN)*, 3rd ed. Cherry Hill, N. J.: Chemical Information Management, 1976.

[74] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Modeling*, vol. 28, no. 1, pp. 31–36, Feb. 1988.

[75] J. Barnard, C. Jochum, and S. Welford, "A universal structure/substructure representation for PC-host communication," in *Chemical Structure Information Systems ; Interfaces, Communication, and Standards*, W. Warr, Ed. Washington, DC: American Chemical Society, 1989, pp. 76–81.

[76] S. Ash, M. a. Cline, R. W. Homer, T. Hurst, and G. B. Smith, "SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation," *Journal of Chemical Information and Modeling*, vol. 37, no. 1, pp. 71–79, Jan. 1997.

[77] S. E. Stein, S. R. Heller, and D. Tchekhovskoi, "An Open Standard For Chemical Structure Representation: The IUPAC Chemical Identifier," in *Proceedings of the International Chemical Information Conference*, 2003, pp. 131–143.

[78] J. M. Barnard, "Structure Representation and Searching," in *Chemical structure systems : computational techniques for representation, searching, and processing of structural information*, J. E. Ash, W. A. Warr, and P. Willett, Eds. New York, New York, USA: Ellis Horwood, 1991, pp. 9–56.

[79] M. F. Lynch and J. D. Holliday, "The Sheffield Generic Structures Project-a Retrospective Review," *Journal of Chemical Information and Modeling*, vol. 36, no. 5, pp. 930–936, Sep. 1996.

[80] J. Klekota, F. P. Roth, and S. L. Schreiber, "Query Chem: a Google-powered web search combining text and chemical structures.," *Bioinformatics (Oxford, England)*, vol. 22, no. 13, pp. 1670–1673, Jul. 2006.

[81] S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Building Chemical Information Systems - the ViFaChem II Project," in *Datenbanksysteme in Business, Technologie und Web (BTW 2009), 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, 2009.

[82] C. L. Borgman, "Social aspects of digital libraries (working session)," in *Proceedings of the first ACM international conference on Digital libraries - DL '96*, 1996, p. 170.

[83] C. L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. The MIT Press, 2007, p. 360.

[84] A. Abdulkader and M. R. Casey, "Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 576–580.

[85] H. Schütze, "Automatic Word Sense Discrimination," *Computational Linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[86] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet And Web Information Systems*, vol. 54, no. 2, pp. 1–17, 1998.

[87] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *Ieee Transactions On Systems Man And Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.

[88] Z. a. Bandar and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871–882, Jul. 2003.

[89] L. Rolling, "Indexing consistency, quality and efficiency," *Information Processing & Management*, vol. 17, no. 2, pp. 69–76, 1981.

[90] F. W. Lancaster, *Indexing and Abstracting in Theory and Practice*. Library Association Publishing, 1991, p. 352.

[91] T. Mann, "'Cataloging Must Change!' and Indexer Consistency Studies: Misreading the Evidence at Our Peril," *Cataloging & Classification Quarterly*, vol. 23, no. 3–4, pp. 3–45, Mar. 1997.

[92] A. Valko and P. Johnson, "CLiDE Pro: A chemical OCR tool," in *Proceedings of the 8th International Conference on Chemical Structures (ICCS)*, 2008.

[93] C. Hurtado, A. Poulovassilis, P. Wood, and S. Spaccapietra, "Query Relaxation in RDF," in *Journal on Data Semantics X*, vol. 4900, S. Spaccapietra, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 31–61.

[94] I. Muslea, "Machine learning for online query relaxation," in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 2004, p. 246.

[95] X. Zhou, J. Gaugaz, W.-T. Balke, and W. Nejdl, "Query relaxation using malleable schemas," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*, 2007, p. 545.

[96] S. Tönnies, B. Köhncke, O. Koepler, and W. Balke, "Exposing the hidden web for chemical digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*, 2010, p. 235.

[97] Z. Hubálek, "Coefficients of Association And Similarity, Based On Binary (Presence-Absence) Data: An Evaluation," *Biological Reviews*, vol. 57, no. 4, pp. 669–689, Nov. 1982.

[98] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical Similarity Searching," *Journal of Chemical Information and Modeling*, vol. 38, no. 6, pp. 983–996, Nov. 1998.

[99] J. Holliday, C. Hu, and P. Willett, "Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings," *Journal of Combinatorial Chemistry; High Throughput Screening*, vol. 5, no. 2, pp. 155–166, 2002.

[100] R. M. Cormack, "A Review of Classification," *Journal of the Royal Statistical Society. Series A (General)*, vol. 134, no. 3, pp. 321–367, 1971.

[101] L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 732–764, 1954.

[102] L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications. II: Further Discussion and References," *Journal of the American Statistical Association*, vol. 54, no. 285, pp. 123–163, 1959.

[103] L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications III: Approximate Sampling Theory," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 310–364, 1963.

[104] P. Willett, "Similarity-based approaches to virtual screening," *Journal of Biochemical Society Transactions*, vol. 31, pp. 603–606, Jun. 2003.

[105] E. Anderson, G. D. Veith, and D. Weininger, "SMILES, a Line Notation and Computerized Interpreter for Chemical Structures," US Environmental Protection Agency, Environmental Research Laboratory, Duluths, 1987.

[106] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen, "Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics," *Current Pharmaceutical Design*, vol. 12, no. 17, pp. 2111–2120, Jun. 2006.

[107] L. H. Hall and L. B. Kier, "Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information," *Journal of Chemical Information and Modeling*, vol. 35, no. 6, pp. 1039–1045, Nov. 1995.

[108] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery," *Journal of Chemical Information and Modeling*, vol. 42, no. 6, pp. 1273–1280, Nov. 2002.

[109] M. G. Kendall, "A New Measure of Rank Correlation," *Journal of Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938.

[110] N. Paskin, "The Digital Object Identifier System," *Encyclopedia of Library and Information Sciences Third Edition*, vol. 3, pp. 1586–1592, 2010.

[111] R. Kimball, *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley, 2004, p. 528.

[112] Y. Cui and J. Widom, "Practical lineage tracing in data warehouses," in *Proceedings of 16th International Conference on Data Engineering*, 2000, pp. 367–378.

[113] J. Widom, "Trio: A System for Integrated Management of Data, Accuracy, and Lineage," in *Proc. of Conference on Innovative Database Systems Research (CIDR)*, 2005, pp. 262–276.

[114] P. Buneman, A. Chapman, and J. Cheney, "Provenance Management in Curated Databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2006, pp. 539–550.

[115] R. Bose and J. Frew, "Composing lineage metadata with XML for custom satellite-derived data products," in *Production*, 2004, vol. 16, pp. 275 – 284.

[116] R. Bose, *A conceptual framework for composing and managing scientific data lineage*. IEEE Comput. Soc, 2002, pp. 15–19.

[117] L. Spery, C. Claramunt, and T. Libourel, "A lineage metadata model for the temporal management of a cadastre application," *Proceedings Tenth International Workshop on Database and Expert Systems Applications DEXA 99*, pp. 466–474, 1999.

[118] P. Missier, K. B. Jjame, J. Zhao, and C. Goble, "Data lineage model for Taverna workflows with lightweight annotation requirements," in *IPAW*, 2008, vol. 5272/2008, pp. 17–30.

[119] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "VisTrails: visualization meets data management," in *SIGMOD Conference*, 2006, vol. 1, pp. 745–747.

[120] I. Altintas, O. Barney, and E. Jaeger-frank, "Provenance Collection Support in the Kepler Scientific Workflow System," *Work*, vol. 4145, pp. 118–132, 2006.

[121] D. Becker, W. McMullen, and K. Hetherington-Young, "A Flexible And Generic Data Quality Metamodel," in *Proceedings of the 12th International Conference on Information Quality*, 2007, pp. 50–64.

[122] A. Copestake, S. Teufel, and B. Waldron, "Flexible Interfaces in the Application of Language Technology to an eScience Corpus," in *Proceedings of the UK e-Science All Hands Meeting (AHM2006)*, 2006.

[123] A. V. Zhdanova, "Community-driven ontology construction in social networking portals," *Journal of Web Intelligence and Agent Systems*, vol. 6, no. 1, pp. 93–121, 2008.

[124] A. Zhdanova, R. Krummenacher, J. Henke, and D. Fensel, "Community-driven ontology management: DERI case study," in *Proc. of the 4th International Conference on Web Intelligence (WI)*, 2005, pp. 73–79.

[125] M. Hatala and G. Richards, "Global vs. community metadata standards: Empowering users for knowledge exchange," *Proc. of the First International Semantic Web Conference*, vol. 2342, pp. 292–306, May 2002.

ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Technische
Universität
Braunschweig